Award Number: W81XWH-13-1-0237

TITLE: IImpact of Noncoding Satellite Repeats on Pancreatic Cancer Metastasis

PRINCIPAL INVESTIGATOR: David T. Ting, MD

CONTRACTING ORGANIZATION: Massachusetts General Hospital
Boston, MA 02114

REPORT DATE: November 2015

TYPE OF REPORT: FINAL

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
November 2015	FINAL	15 Aug 2013 - 14 Aug 2015

4. TITLE AND SUBTITLE

Impact of Noncoding Satellite Repeats on Pancreatic Cancer Metastasis

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-13-1-0237

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David Ting, MD
Daniel A. Haber. M.D. Ph.D. (Mentor)

E-Mail: dting1@partners.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts General Hospital, The
55 Fruit Street
Boston, MA 02114-2621

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

*State the purpose, scope, major findings and be an **up-to-date** report of the progress in terms of results and significance*. (Approx. 200 words)

The goal of the project is to understand the role of HSATII satellite repeat expression in pancreatic cancer metastatic potential. An inducible over-expression vector was created and was successfully used in cancer cell lines with evidence of transcriptional changes and increased migratory capability consistent with a role in metastasis. HSATII was also assessed in pancreatic circulating tumor cells (CTCs), which are enriched for metastatic precursors. Initial results find these cells in patients with preneoplastic IPMN lesions suggesting a blood based early detection biomarker. Unexpectedly, a novel reverse transcriptional machinery has been identified with HSATII expression and this results in genomic expansion of these pericentromeric repeats in cancer. This has provided a new understanding of HSATII regulation and function, which has adjusted the goals of the project to address these new findings.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	36	19b. TELEPHONE NUMBER *(include area code)*

TABLE OF CONTENTS

1. INTRODUCTION

Pancreatic cancer remains one of the most deadly cancers where the vast majority of patients are diagnosed too late and conventional therapies have largely been ineffective, making early detection and novel drug targets greatly needed. Recent studies have shown the expression of a significant portion of genomic regions previously thought to be transcriptionally silent. Satellites are regions of the genome that are highly repetitive and normally their expression is suppressed by heterochromatin, however, their expression was found to be abundant in a wide variety of cancers. In particular, the HSATII satellite was found to be specifically upregulated in cancer cells compared to all other repetitive elements and was found to be highly elevated in preneoplastic lesions of the pancreas, which has implications as a novel early detection biomarker. Moreover, the ability to target HSATII in cancers may offer a novel therapeutic avenue. The goal of this research is to understand the cellular and molecular impact of satellite RNA in cancer cells and to test the utility of these highly specific and abundant transcripts as novel biomarkers for early detection.

2. KEYWORDS

cancer genetics, satellite repeats, metastasis, circulating tumor cell, pancreatic cancer

3. ACCOMPLISHMENTS

Aim 1: Evaluation of Satellite expression on transcriptional profiles

Task 1. Development of satellite expressing cell lines: A doxycycline inducible HSATII vector was created containing a segment of HSATII that was approximately 800 bp in length. This vector was successfully transduced into human cancer cell line SW620. Induction of HSATII expression was clearly evident by the addition of doxycycline and confirmed by RNA-ISH (Fig. 1) and northern blot.

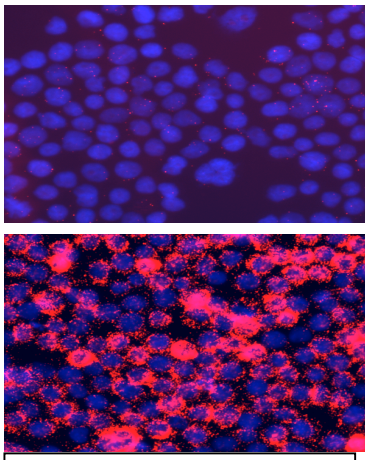


Fig. 1: SW620 HSATII inducible cell line without (top) and with doxycycline (bottom). HSATII RNA-ISH (red) and DAPI nuclear stain (blue)

Task 2. Effects of satellite on expression patterns: Satellite induction was performed with doxycycline and cell lines were evaluated for expression pattern changes using the Helicos single molecule sequencing platform. A total of 126 genes were differentially expressed in HSATII induced cell lines compared to GFP induced cell line controls. Gene ontology of these genes did not identify any major signaling pathways or other known expression profile enrichment. Notably, 46% of these genes are upregulated in human brain tissue consistent with our prior correlation of satellite expression to neural genes.

Although forced over-expression of HSATII did result in transcriptional and functional changes, we had learned of a way to induce endogenous HSATII expression, which we believe was more physiologic. We modified our statement of work to reflect this change.

Our previous work, had shown that although HSATII was highly expressed in tumors, upon culturing cancer cells in standard 2D adherent culture there is complete suppression of HSATII (Fig. 2). Because of this finding, we amended our SOW to include a new task 3.

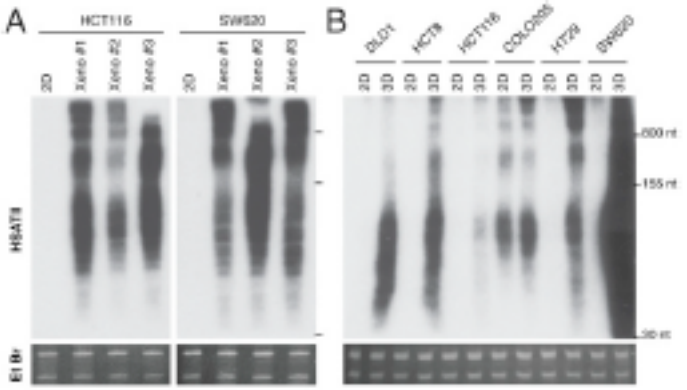


Fig. 2: HSATII induction from growth as tumors or tumorspheres. Northern blot of A) HCT116 and SW620 cancer cell lines grown in 2D and Xenografts. B) Panel of cancer cell lines grown in 2D or 3D

Task 3: Understanding satellite deregulation:
We refocused this task to understand satellite deregulation in cell lines. We performed a number of perturbations in vitro to induce HSATII expression including hypoxia, UV radiation, demethylation, starvation, and growth in non-adherent conditions. Only growth in non-adherent conditions (as 3D tumor spheres or in soft agar) was sufficient to induce HSATII expression in multiple cancer cell lines including pancreatic and colon

cancers (Fig. 3). We have completed RNA-seq analysis of 2D and 3D cells paired RNA-seq and are continuing to analyze this data. We initially planned on performing targeted ChIP-seq, but are awaiting complete analysis of the RNA-seq data before pursuing this line of work. In summary, we have found an interesting phenomenon where HSATII expression is dynamically controlled by growth in adherent/non-adherent conditions. This allows for a model tool to study the regulation of HSATII and also offers a means to understand the impact of endogenous HSATII expression in cancer cells.

Aim 2: Evaluation of Satellite expression on metastatic potential

Task 1. Effects on adherent culture: Effects of HSATII induction in HSATII cell lines was performed showing some changes in morphology of the cell line and increased migratory function. However, there were no appreciable effects on cell growth and a potentially negative effect on tumor sphere formation with HSATII overexpression. Our work on Aim 1 Task 3 led us to see that HSATII induction through growth in non-adherent conditions led to expression of multiple transcript sizes and not a single size as we have done through our vector.

When thinking about the potential functions of HSATII RNA, we began to compare HSATII to other repeats in the genome. Previously, we had found satellites to be highly co-expressed with the LINE-1 element across mouse and human cancers. LINE-1 is an active retrotransposon that inserts throughout the euchromatin of genomes and recent publications have shown that LINE-1 retrotransposition is a common event across a wide variety of malignancies. In addition, the telomere repeat is also reverse transcribed through telomerase in the vast majority of cancers. The parallels of these two major repeats led us to hypothesize that HSATII may be reverse transcribed (RT) as a means to expand these regions in tumor genomes. We evaluated the presence of HSATII RT products by treating xenograft small nucleic acid extracts with DNase I and indeed found sensitivity of these species, which indicated the presence of dsDNA (Fig. 3). Because of this unexpected finding, we focused our efforts in validating this novel finding and to understand the significance of this phenomenon in cancer function. Since LINE1 retrotransposition activity in colon cancer had been previously been shown, we used colon cancer as a model to best study this novel reverse transcriptional mechanism. Through a number of biochemical experiments we believe reverse transcriptional machinery is highly active and specific for satellite repeats in human cells. These RNA derived DNAs (rdDNA) are found in primary tumors, xenografts, and tumorspheres in large amounts and appear to be used as templates for elongating the pericentromeric regions from where they originate. We validated the DNA expansion of HSATII regions in our xenograft models and find 50% of primary colon cancers with significant copy number gains of HSATII as determined by whole genome sequencing. Importantly, these HSATII copy number gains were found to confer a worse prognosis in these cancers (Fig. 4). Interestingly, we found that certain HIV RT inhibitors (ddC) could block the generation of HSATII DNA/RNA hybrids, which had anti-cancer effects in 3D and in xenografts (see Task 2). **See attached publication in press at PNAS for more details**.

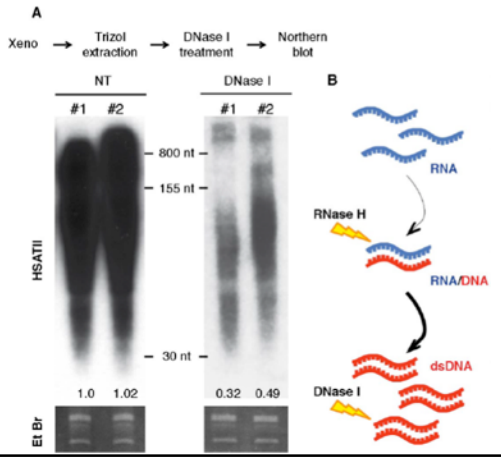


Fig. 3: A) Northern blot analysis for HSATII in colon cancer xenografts treated with DNase I showing partial digestion. B) Model of HSATII RT and predicted sensitivity to nucleases.

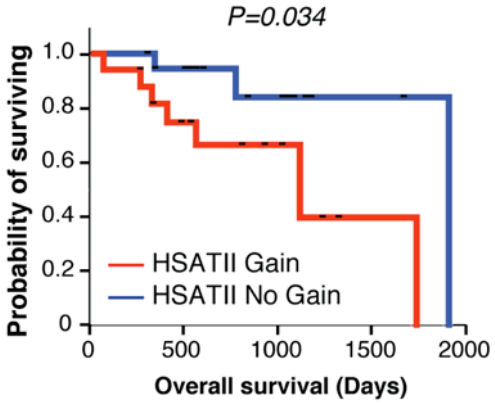


Fig. 4: Kaplan Meier survival analysis of colon cancer TCGA data. HSATII copy number gain (red) and no gain (blue). Log-rank p-value shown.

Task 2. Effects on xenograft tumors: Due to the findings of differential expression of HSATII in 2D culture compared to 3D and xenografts, we focused on these differences as a means to understand HSATII function.

Given the gains of HSATII copy number from the RT process, we hypothesized that blockage of HSATII RT may have differential effects in 3D and xenografts compared to 2D culture. Indeed, we find that using the RT inhibitor ddC or using HSATII sequence specific locked nucleic acids did have an anti-cancer effect both in 3D tumor spheres and in xenografts (Fig. 5). Together this has revealed a new therapeutic vulnerability in cancer that can potentially be rapidly translated to the clinic given the utility of established drugs for HIV.

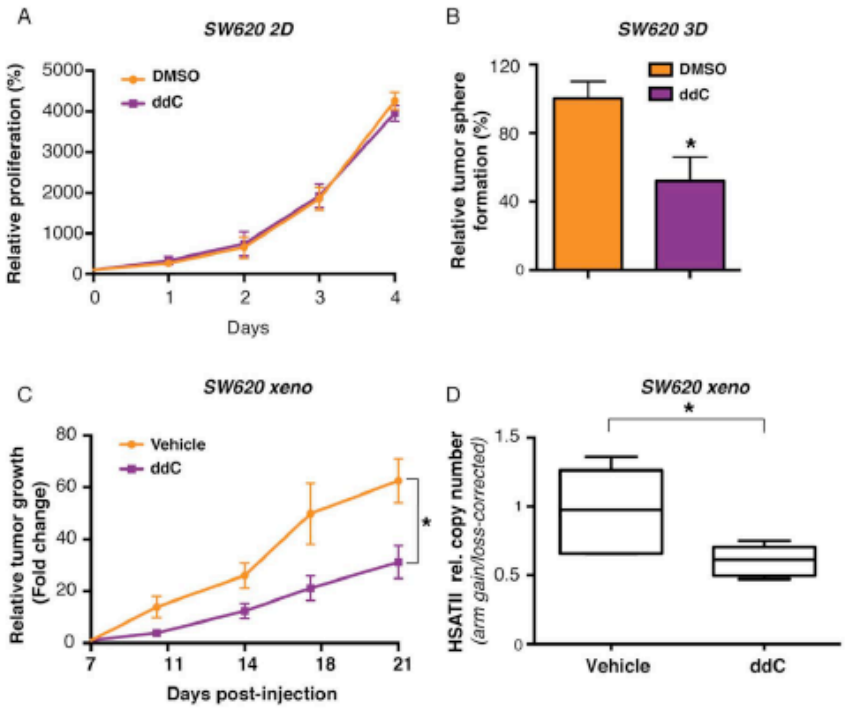


Fig. 5: NRTI ddC differential effect in SW620 cell line grown in (A) 2D vs (B) 3D. NRTI ddC with significant reduction in SW620 xenograft (C) proliferation and (D) genomic HSATII copy number gain

Task 3. Effects on CTCs: Due to our findings above we have deferred looking at the effects of HSATII on CTCs. However, we note that the disruption of HSATII RT affects tumor sphere formation, which we had previously shown to be a critical surrogate functional assay for CTC viability (Yu M*, Ting DT*, et al. Nature 2012). Recent, analysis of mouse pancreatic CTC data has identified elevated repeat expression in CTCs compared to matched primary tumors pointing towards a correlation of increased satellite expression in CTCs compared to primary tumor cells. Human HSATII expression in CTCs (See Aim 3) has also shown increased detection sensitivity of CTCs again pointing towards a relationship of satellite expression to the metastatic process. **We plan to formally evaluate effects on CTCs in the future using the pancreatic genetically engineered mouse model, which was outside the scope of this grant.**

Aim 3: Evaluating Satellites as a novel CTC Biomarker

Task 1. Optimization of RNA-ISH assay for HSATII in CTCs: HRPO approval and initiation of testing HSATII ISH in clinical samples has been done over the last year. Initial testing of RNA-ISH on the 3rd generation IFD CTC-chip has been completed. Automated imaging analysis has been optimized and standard operating procedures have been established for the CTC diagnostic.

Task 2. Comparative analysis of HSATII RNA-ISH versus CK/EpCAM Immunofluorescence CTC enumeration assays:
We have initiated a comparison of the standard CK immunofluorescence (IF) to HSATII ISH in a split blood sample run on the IFD CTC-chip in both resectable pancreatic cancer and patients with cystic lesions of the pancreas called intraductal papillary mucinous neoplasms (IPMN). IPMNs are often found incidentally and only approximately 5-10% will progress to invasive cancer. Understanding which patients require definitive surgical resection and ones that can be monitored is of

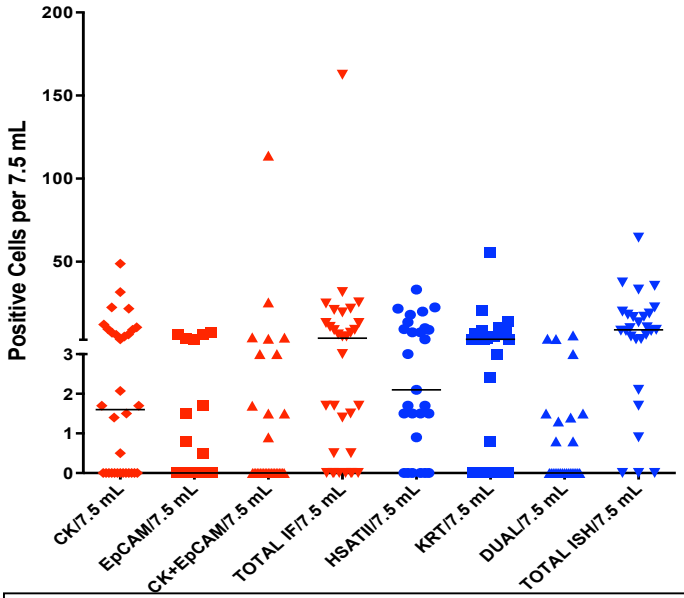


Fig. 6: CTCs identified by CK and EpCAM IF (Red) or HSATII and KRT RNA-ISH (Blue) from IPMN patients

great importance for the field. CTCs may provide a means to predict which IPMN patients may be at risk of invasive disease. CTC counts with standard CK+EpCAM compared to HSATII+KRT RNA-ISH show comparable sensitivity in IPMN patients tested (Fig. 6). Both IF and RNA-ISH performed well in this pilot with RNA-ISH demonstrating a higher sensitivity of 74% compared to 50% using IF. We are continuing to monitor these patients for the development of pancreatic cancer to see which marker or combination of markers is most prognostic in this at-risk patient population. In parallel, we have also completed RNA-sequencing of CTC samples from patients with IPMN and resectable pancreatic cancer to define the transcriptional changes (coding and non-coding) that occur between preneoplastic and invasive carcinoma CTCs.

What opportunities for training and professional development has the project provided?

Nothing to Report

How were the results disseminated to communities of interest?

Nothing to Report

What do you plan to do during the next reporting period to accomplish the goals?

Nothing to Report

4. IMPACT

The massive expression of satellite repeats in virtually all epithelial cancers was an unexpected finding with implications as a cancer diagnostic and also as a new unappreciated phenomenon in cancer biology. In our pursuit, to understand the biological regulation and function of satellites, we have discovered a novel reverse transcriptional mechanism that expands satellite repeat regions in the genome, which when blocked leads to anti-cancer effects in both tumor spheres and in xenografts. Increased levels of satellites in circulating tumor cells (CTCs) supports a relationship of HSATII with metastasis and more importantly this may prove to be an blood based early detection diagnostic for pancreatic cancer.

This work has led to fruitful collaborations with others as well as additional funding. A collaboration with Benjamin Greenbaum and Nina Bhardwaj at MT. Sinai Cancer Center has led to the discovery that HSATII RNA is immunostimulatory to human macrophages (See attached manuscript Tanne et al. PNAS 2015). This work combined with the work we have done indicates that HSATII RNA not only has tumor cell specific, but also microenvironmental effects that are important in cancer cell progression. This has led to the formation of a SU2C-NSF-V Foundation Convergence team focused on understanding the role of immunotherapy in pancreatic cancer. As part of this collaborative research group, I will be studying the role of HSATII RNA in this context. In summary, we have made significant progress in understanding the mechanistic underpinnings of these repetitive elements in cancer, identified a novel therapeutic opportunity in disrupting these repeats, and developed a new blood based diagnostic platform based on HSATII RNA.

5. CHANGES/PROBLEMS

All changes for this project are detailed in section 3. Accomplishments. These were approved by the grants administrator and SOW was modified to reflect these changes.

6. PRODUCTS

Publications (related)

1. Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H, Ramaswamy S, Park PJ, Maheswaran S, **Ting DT#**, Haber DA#. Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *PNAS*, (2015). published ahead of print

November 2, 2015, doi:10.1073/pnas.1518008112. #Co-corresponding. Published. Acknowledgement of Federal Support (YES)

2. Tanne A, Muniq LR, Puzio-Kuter A, Leonova KI, Gudkov AV, Ting DT, Monasson R, Cocco S, Levine AJ, Bhardwaj N, Greenbaum BD. Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *PNAS*, (2015). published ahead of print November 2, 2015, doi:10.1073/pnas.1517584112. Published. Acknowledgement of Federal Support (YES)

Oral Presentations

1. **Ting DT**. Pancreatic Circulating Tumor Cells: Window into the Metastatic Cascade. San Diego, CA. Nov. 2015

2. **Ting DT**. Pancreatic Cancer Liquid Biopsies. Cold Spring Harbor, NY. June 2015

3. **Ting DT.** Single cell RNA-sequencing of Circulating Tumor Cells. AACR 2015. Philadelphia, PA April 2015

4. **Ting DT.** Single Cell RNA-Sequencing of Pancreatic Circulating Tumor Cells. Gordon Research Conference. Mt. Holyoke, MA. Aug 2014

5. **Ting DT.** Diversity of Circulating Tumor Cells in a Mouse Pancreatic Cancer Model Identified by Single Cell RNA Sequencing. AACR 2014. San Diego, CA. April 2014

Inventions, Patents and Licenses

Targeting Human Satellite II (HSATII): U.S. Patent Application Serial No. 62/017,012, filed on June 25, 2014 Attorney Docket No. 29539-0125WO1; MGH 22846 (Provisional)

Reportable Outcomes

Nothing to report

Other Achievements

Nothing to report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

Name:	*David Ting*
Project Role:	*PI*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*2.4*
Contribution to Project:	*Dr. Ting had overall responsibility for the project and contributed with the design and performance of experiments.*
Funding Support:	This award
Name:	Olivia MacKenzie
Project Role:	Research Technician
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	12
Contribution to Project:	Ms. MacKenzie provided technical support for the RNA-ISH and CTC aspects of the project.
Funding Support:	This award

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Active (Changes for duration of Award)

Translational Continuation Research Grant Pancreatic Cancer Action Network	Ting (PI)	7/1/14-6/30/16

Circulating Tumor Cells to Assess Pancreatic Cancer Disease Status
The goal of this project is to assess the potential of pancreatic CTCs as a prognostic and predictive biomarker for pancreatic cancer in a clinical trial of patients with resectable disease.
Role: PI

Research Scholar Grant Pandharipande (PI) 7/1/17-6/30/19
American Cancer Society
Pancreatic Cancer Screening: Development of a Simulation Model
The goal of the proposed research is to reduce the burden of pancreatic cancer by identifying effective programs for early detection
Role: Co-Investigator

Completed (Changes)
Warshaw Institute; Mass General Hospital Dept of Surgery Ting (PI) 7/1/14-6/30/15
Andrew L. Warshaw, M.D. Institute for Pancreatic Cancer Research Fellowship
Role of Cancer Associated Fibroblasts in Pancreatic Cancer Progression
The goal of this project is to understand the contribution of cancer associated fibroblasts in tumor growth and dissemination.
Role: PI

What other organizations were involved as partners?

There was no significant collaboration directly linked to the work. However, a fruitful collaboration with Drs. Greenbaum and Bhardwaj at the Mt. Sinai Cancer Center did occur as a product of this research project. This led to assistance for their work on the immunostimulatory properties of HSATII that led to a publication for which I am a co-author (attached in Appendix).

8. SPECIAL REPORTING REQUIREMENTS

N/A

9. APPENDICES

See attached manuscripts that have been published online at PNAS Early Edition and will follow in print in the next month.

Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer

Francesca Bersani[a], Eunjung Lee[b,c], Peter V. Kharchenko[b,d], Andrew W. Xu[b], Mingzhu Liu[a,e], Kristina Xega[a], Olivia C. MacKenzie[a], Brian W. Brannigan[a], Ben S. Wittner[a], Hyunchul Jung[f], Sridhar Ramaswamy[a,g], Peter J. Park[b,c,h], Shyamala Maheswaran[a,i], David T. Ting[a,g,1,2], and Daniel A. Haber[a,e,g,1,2]

[a]Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown, MA 02129; [b]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115; [c]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115; [d]Hematology/Oncology Program, Children's Hospital, Boston, MA 02115; [e]Howard Hughes Medical Institute, Chevy Chase, MD 20815; [f]Samsung Medical Center, Seoul 135-710, Korea; [g]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114; [h]Informatics Program, Children's Hospital, Boston, MA 02115; and [i]Department of Surgery, Massachusetts General Hospital, Boston, MA 02114

Edited by Carol Prives, Columbia University, New York, NY, and approved October 5, 2015 (received for review September 11, 2015)

Aberrant transcription of the pericentromeric human satellite II (HSATII) repeat is present in a wide variety of epithelial cancers. In deriving experimental systems to study its deregulation, we observed that HSATII expression is induced in colon cancer cells cultured as xenografts or under nonadherent conditions in vitro, but it is rapidly lost in standard 2D cultures. Unexpectedly, physiological induction of endogenous HSATII RNA, as well as introduction of synthetic HSATII transcripts, generated cDNA intermediates in the form of DNA/RNA hybrids. Single molecule sequencing of tumor xenografts showed that HSATII RNA-derived DNA (rdDNA) molecules are stably incorporated within pericentromeric loci. Suppression of RT activity using small molecule inhibitors reduced HSATII copy gain. Analysis of whole-genome sequencing data revealed that HSATII copy number gain is a common feature in primary human colon tumors and is associated with a lower overall survival. Together, our observations suggest that cancer-associated derepression of specific repetitive sequences can promote their RNA-driven genomic expansion, with potential implications on pericentromeric architecture.

satellites | reverse transcription | repeats | cancer

Pericentromeric satellite repeats are essential core centromere-building elements that stabilize interactions with DNA-binding proteins, maintain heterochromatin architecture, sustain kinetochore formation, and drive chromosomal segregation during mitosis, thereby ensuring faithful duplication of the genome (1). Transcription from pericentromeric satellites has been reported in plants and invertebrates, as well as during early stages of vertebrate development, and some types of satellite repeats are induced following environmental stress in cell line models (2). We recently reported the massive overexpression of specific classes of satellite repeats in human epithelial cancers, resulting from aberrant transcription of these pericentromeric domains (3). In almost all cancers analyzed, subsets of pericentromeric satellites are expressed at very high levels (3–5), whereas others show consistently reduced expression compared with normal tissues.

The human satellite II (HSATII) is the most differentially expressed satellite subfamily in epithelial cancers (3). It constitutes the main component of pericentromeric heterochromatin on chromosomes 2, 7, 10, 16, and 22 (UCSC Genome Browser, genome.ucsc.edu), and it is also found at chromosome band 1q12, where it is colocated with satellite III sequences. HSATII is defined by tandemly repeated divergent variants of 23- to 26-bp consensus sequences, organized in long arrays that may span up to thousands of kilobases (6). Although repetitive DNA sequences are frequently hypomethylated in cancer cells (7), the mechanisms underlying their aberrant expression are not well understood. For instance, loss of DNA methylation alone does not result in overexpression of the HSATII satellite (8), suggesting the existence of more complex regulatory networks.

In establishing models to study the molecular basis of HSATII RNA overexpression in cancer, we found that growth of cells under nonadherent conditions is sufficient to trigger induction of this satellite repeat. Unexpectedly, under these and other conditions, we uncovered that these repeated transcripts are reverse-transcribed into DNA/RNA hybrids. The reintegration of HSATII RNA-derived DNA (rdDNA) is correlated with a progressive expansion of host HSATII genomic loci. These results point to an unexpected plasticity of pericentromeric repeat-containing structures during cancer progression.

Results and Discussion

Detection of HSATII Expression in Human Tumors and 3D Cancer Cell Models. The highly repetitive nature of satellites precludes their precise quantitation and qualitative analysis using PCR-based RNA sequencing approaches. We previously showed that PCR-independent single molecule next-generation sequencing [digital gene expression (DGE) profiling] is uniquely sensitive and quantitative (9), although it is not suited to routine analysis and does not provide qualitative length information. To enable experimental models for the study of HSATII deregulation, we first designed a modified Northern blot HSATII assay (*SI Appendix*, Fig. S1*A*). HSATII satellite transcripts encompass arrays of variable lengths derived from multiple different genomic locations (6); thus, Northern blotting generates a pattern of bands ranging from 30 nt to greater than 800 nt in size (*SI Appendix*, Fig. S1*B*), consistent with the pattern reported for other repeats,

Significance

Unique among the large number of noncoding RNA species, the pericentromeric human satellite II (HSATII) repeat is massively expressed in a broad set of epithelial cancers but is nearly undetectable in normal tissues. Here, we show that deregulation of HSATII expression is tightly linked to growth under nonadherent conditions, and we uncover an unexpected mechanism by which HSATII RNA-derived DNA (rdDNA) leads to progressive elongation of pericentromeric regions in tumors. The remarkable specificity of HSATII overexpression in cancers, together with the consequences of targeting its RT, points to a potential novel vulnerability of cancer cells.

Author contributions: F.B., S.M., D.T.T., and D.A.H. designed research; F.B., K.X., O.C.M., and B.W.B performed research; F.B., E.L., P.V.K, A.W.X., M.L., B.S.W., H.J., S.R., and P.J.P. contributed new reagents/analytic tools; F.B. and D.T.T. analyzed data; and F.B., S.M., D.T.T., and D.A.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]D.T.T. and D.A.H. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: dting1@mgh.harvard.edu or dhaber@mgh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518008112/-/DCSupplemental.

such as murine minor satellites (10) and human satellite III (11). Quantitation of DGE profiles and Northern blot signal intensity for matched primary cancer specimens were highly correlated (*SI Appendix*, Fig. S1 *B* and *C*).

We observed that human colorectal cancer cell lines do not express HSATII under standard in vitro adherent (2D) culture conditions, but strongly up-regulate its expression when grown as tumor xenografts (Fig. 1*A*). To define specific experimental conditions that modulate HSATII expression within tumors, we tested multiple stimuli associated with cellular stress and tumorigenesis, including hypoxia, UV irradiation, heat shock, oxidative stress, overconfluence, treatment with demethylating agents, coculturing with stromal-derived feeder layers, and culture under anchorage-free conditions (*SI Appendix*, Fig. S2 *A–D*). Remarkably, only culture under nonadherent conditions, as 3D tumor spheres in solution or in soft agar, led to robust induction of HSATII in five colorectal cancer cell lines, as detected by Northern blotting (Fig. 1*B*). The specific induction of HSATII RNA by anchorage-independent growth was also demonstrated using RNA in situ hybridization, an imaging assay that does not involve nucleic acid extraction or denaturation, confirming the RNA specificity of the hybridization signal (Fig. 1*C*). A sixth colorectal cancer line, COLO205, noteworthy for its semiadherent growth pattern, was unique in expressing HSATII RNA at baseline under standard culture conditions (Fig. 1*B*). Notably, the elevated RNA levels detected in tumor spheres and xenografts from two independent cell lines were rapidly lost upon reestablishing the growth of these cells under adherent 2D culture (*SI Appendix*, Fig. S2*E*). Consistent with our previous findings in primary tumors (3), HSATII transcripts were present in both sense (S) and antisense (AS) orientations (*SI Appendix*, Fig. S2*F*), and they were primarily localized to the nuclear compartment in a panel of colon cancer cell lines (*SI Appendix*, Fig. S2*G*), similar to other repetitive noncoding RNAs such as telomeric repeat-containing RNA (TERRA) (12).

Previous studies have shown that satellite expression in both human and mouse cells may be due to a combination of environmental stimuli and genetic factors. For instance, human satellite III is expressed in response to cellular stress, including UV-C, oxidative, heat shock, and hyperosmotic stress (13), whereas mouse pericentromeric major satellites can be induced by growth to confluence or treatment with hypomethylating, apoptotic, or differentiation-inducing agents (10). Genetic lesions in the tumor suppressor *BRCA1* lead to Alpha-satellite sequence derepression in human breast epithelial cells (5), and deletion of *Trp53* in mouse cells results in murine major satellite expression (14). Remarkably, the HSATII pericentromeric repeat is resistant to general environmental stressors and differences in oncogene or tumor suppressor genotypes under standard adherent culture. Only growth under nonadherent conditions is sufficient to induce robust HSATII expression across different cancer cell lines, an interesting phenomenon that merits further investigation.

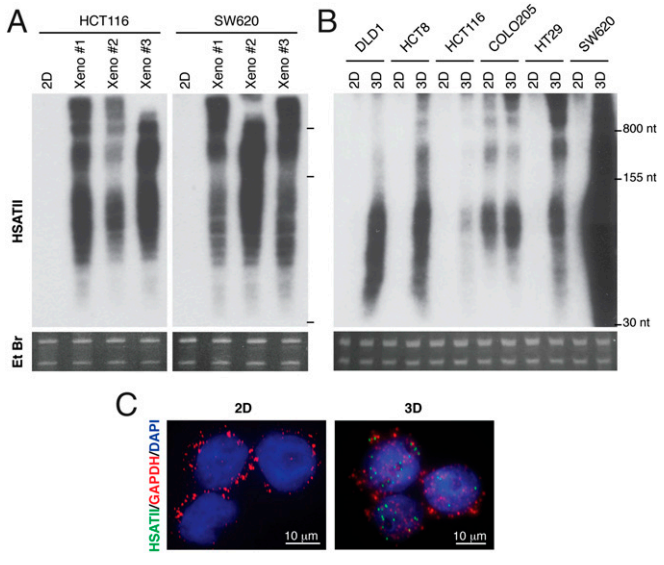


Fig. 1. HSATII is expressed in human tumors and 3D cancer cell models. (*A*) Northern blot analysis of HSATII expression in HCT116 and SW620 cells grown as 2D cultures or xenografts (Xeno). (*B*) Northern blot analysis of HSATII expression in colon cancer cell lines grown as 2D cultures or tumor spheres (3D). Ethidium bromide (Et Br) stainings of gels are shown for each Northern blot analysis as loading controls. (*C*) RNA in situ hybridization (with the indicated fluorescent probes) of SW620 cells cultured under 2D conditions or as tumor spheres (3D). HSATII/DAPI colocalization coefficient measured by confocal imaging: $R = 0.6 \pm 0.04$.

Identification of Medium/Small-Molecular-Weight HSATII DNA. Unexpectedly, we observed that a fraction of the xenograft-induced HSATII sequences present within medium/small-molecular-weight nucleic acids (extraction with Trizol; ThermoFisher) was sensitive to DNase I (Fig. 2*A*). This event may result from genomic DNA (gDNA) contamination or from the existence of small dsDNA, dsRNA partially susceptible to DNase I, and/or DNA/RNA hybrids. To exclude the possibility of gDNA contamination, we performed multiple controls. First, during nucleic acid extraction, we applied a solid barrier to prevent any cross-contamination between the aqueous phase and the organic, gDNA-containing phase (phase lock gel tubes) (*SI Appendix*, Fig. S3*A*). Second, we showed that high-molecular-weight HSATII gDNA from a colon cancer xenograft is readily distinguished by Northern blotting from separately processed small-molecular-weight RNA/DNA from the same tissue, without evidence of cross-contamination (*SI Appendix*, Fig. S3*B*). Third, using multiple cell lines, we showed that the presence of medium/small-molecular-weight HSATII sequences is only evident following culture under 3D or xenograft conditions. Identical and simultaneous analysis of these cells cultured under 2D conditions showed no evidence of these small HSATII sequences (Fig. 1 *A* and *B* and *SI Appendix*, Fig. S2). Finally, under identical processing conditions, the rapid loss of HSATII signal following replating of 3D cultures into 2D cultures further excluded gDNA contamination as a source for the HSATII DNA signal on Northern blotting (*SI Appendix*, Fig. S2*E*). We therefore concluded that deregulated HSATII satellite transcripts coexist with matched DNA fragments within the medium/small-molecular-weight nucleic acid fraction of cells that overexpress this satellite repeat. Given the presence of small RNA and DNA species, we hypothesized that HSATII RNA may be reverse-transcribed into DNA/RNA hybrids and, ultimately, dsDNA (*SI Appendix*, Fig. S3*C*), a phenomenon known to occur with other repetitive elements, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeat (LTR) retrotransposons (15).

Generation of DNA/RNA Hybrids upon Ectopic Introduction of HSATII RNA. To capture the HSATII RNA-to-DNA conversion, we first developed an assay to introduce synthetically produced HSATII RNA generated by in vitro transcription (IVT) directly into 2D-cultured 293T cells that lack endogenous HSATII expression (*SI Appendix*, Fig. S3*D*). To assess the formation of DNA/RNA hybrids in cells transfected with single-stranded IVT HSATII RNA (HSATII-chr10), we subjected nucleic acid extracts to treatment with RNase H, which specifically digests the RNA moiety of DNA/RNA hybrids but does not affect either ssRNA or the DNA component of DNA/RNA hybrids (Fig. 2*B*). Indeed, RNase H treatment caused a strong reduction in the Northern blot signal identified for the HSATII S (but not AS) sequence (Fig. 2*C* and *SI Appendix*, Fig. S3*E*). Thus, consistent with the generation of rdDNA, a fraction of the IVT-produced RNA is within a complex with a cDNA strand. Transfection of comparable amounts of IVT GFP RNA produced the expected RNA signal but showed no significant sensitivity to RNase H (Fig. 2*D*),

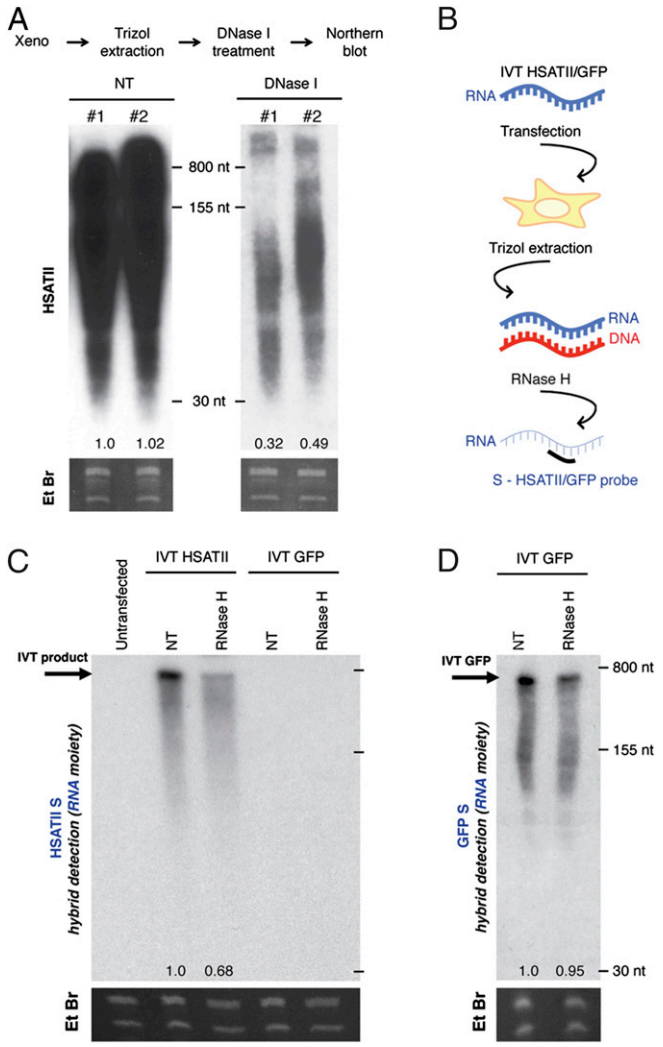


Fig. 2. Ectopic HSATII RNA gives rise to DNA/RNA intermediates. (*A*) Northern blot analysis of HSATII in untreated (NT) or DNase I-treated extracts obtained from SW620 xenografts. Numbers below indicate relative signal quantitation. (*B*) TRIzol extracts obtained from 293T cells 24 h after transfection with HSATII/GFP IVT products were subjected to RNase H treatment, followed by Northern blotting and hybridization to detect the RNA strand (S-HSATII) of the hybrid. (*C*) Northern blot analysis of extracts from 293T cells either untransfected or transfected with IVT HSATII or GFP, subjected to the indicated nuclease treatment and probed for HSATII S. (*D*) Northern blot analysis of extracts from 293T cells after transfection with IVT GFP, treated with RNase H and probed for GFP S. Numbers below indicate relative signal quantitation.

and introduction of GFP RNA did not lead to the induction of HSATII satellite RNA (Fig. 2*C*, last two lanes). These findings exclude the occurrence of nonspecific responses to RNA transfection. We conclude that introduction of purified S HSATII RNA generates a DNA/RNA hybrid in IVT RNA-transfected cells. Because IVT using T7 polymerase relies on a PCR-generated DNA template as starting material, we included multiple controls to ensure the absence of DNA template contamination within the IVT product itself, as well as any genomic HSATII sequences in the cellular extracts (*SI Appendix*, Fig. S3 *F–L*). These results indicate that ectopically introduced single-stranded HSATII RNA is capable of generating cDNA within transfected cells.

To validate these results further, we used the S9.6 monoclonal antibody, which is highly specific in its recognition of DNA/RNA hybrids (16–19). We established a DNA/RNA hybrid immunoprecipitation (DRIP) assay using quantitative PCR (qPCR) of S9.6 immunoprecipitates (HSATII-chr10 qPCR), which was applied to nucleic acids from untransfected or IVT HSATII RNA-transfected cells. Samples were first subjected to complete DNase I digestion (which removes all dsDNA but does not affect DNA/RNA hybrids), followed by DRIP analysis (Fig. 3*A*, *Left*). HSATII DNA/RNA duplexes were present only in 293T cells transfected with HSATII RNA, and treatment of extracts with RNase H effectively abolished immunoprecipitation of the HSATII DNA/RNA hybrids (Fig. 3*B*). Taken together, the formation of HSATII RNA/DNA hybrids in a controlled IVT model and the detection of these species by DRIP are suggestive of RT.

RT of Endogenous HSATII RNA into DNA/RNA Hybrids. To extend these analyses to a more physiological context, we assessed the presence of endogenous HSATII DNA/RNA hybrids by applying the DRIP assay to SW620 colon cancer cells grown as 3D tumor spheres (Fig. 3*A*, *Right*). RNase H-sensitive DNA/RNA HSATII hybrids were immunoprecipitated using the DRIP assay in SW620 spheroids (Fig. 3*C*). DNA/RNA hybrids were also identified in COLO205 semi-attached cell cultures (*SI Appendix*, Fig. S3*M*), which are characterized by baseline expression of HSATII transcripts (Fig. 1*B*).

To evaluate the consequences of RT inhibition on the formation of these hybrids, we tested the effect of the nucleoside analog reverse transcriptase inhibitor (NRTI) 2′,3′-dideoxycytidine (ddC) in HSATII-expressing cells (Fig. 3*A*, *Right*). Notably, ddC is very poorly incorporated by replicative polymerases (20, 21), although it displays high specificity for multiple classes of RT, including LINE-1 (22). Treatment of SW620 spheroids and COLO205 cells with ddC significantly reduced the levels of endogenous HSATII DNA/RNA hybrids, as measured by the DRIP assay (Fig. 3*C* and *SI Appendix*, Fig. S3*M*). These observations are consistent with RT activity in HSATII-expressing cells, contributing to the generation of DNA/RNA structures derived from satellite transcripts.

RT activity in mammalian cells is derived primarily from retrotransposons, including LINE-1, the human endogenous

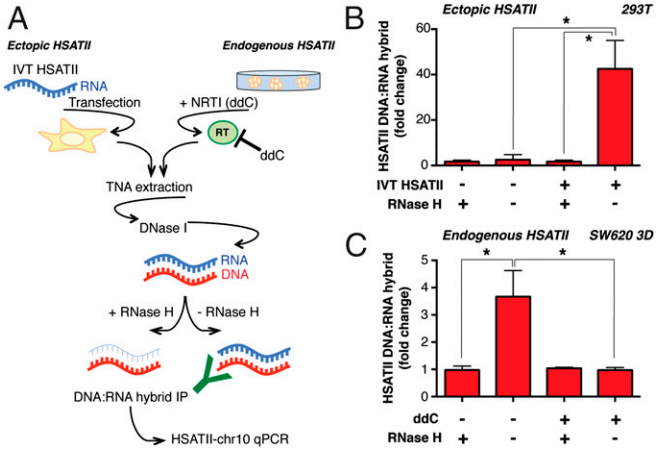


Fig. 3. DRIP reveals the presence of ectopic as well as endogenous HSATII hybrids whose production is affected by RT inhibition. (*A*) Outline of the experimental layout. Total nucleic acids (TNAs) were isolated from IVT HSATII-transfected 293T cells or SW620 tumor spheres cultured in the presence of ddC or DMSO (ddC⁻). TNAs were treated with DNase I digestion to remove all potential gDNA contamination. DNA/RNA hybrids were then purified by immunomagnetic pull-down using a hybrid-specific antibody, and their relative quantities were measured by HSATII-chr10 qPCR. Pretreatment of TNA samples with RNase H as indicated demonstrates abrogation of DNA/RNA hybrid detection. (*B*) Fold change in the enrichment of DNA/RNA hybrids in HSATII-transfected 293T cells measured by qPCR after DRIP. (*C*) Fold enrichment of endogenous HSATII DNA/RNA hybrids in SW620 tumor spheres analyzed by HSATII-chr10 qPCR after DRIP. Fold changes were calculated based on percent input values, and the RNase H-treated samples were set at 1. For all charts, values represent the average of three independent experiments ± SEM. *$P < 0.05$ (*t* test).

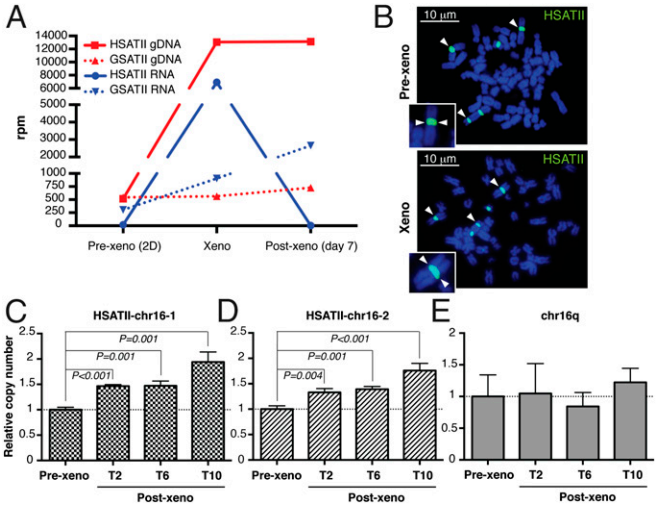


Fig. 4. HSATII rdDNA is reintegrated at the same original loci in the genome, leading to pericentromere elongation in colon cancer xenografts. (*A*) DGE (RNA) and copy number (gDNA) analysis of satellite repeats (HSATII and GSATII) in the indicated samples (SW620) quantitated by single molecule sequencing. (*B*) Representative HSATII DNA FISH (white arrowheads) on metaphase spreads of prexenograft (Pre-xeno) 2D cultures and xenografts obtained from SW620 cells. (*Insets*) Enlarged (1,000×) HSATII-positive chromosomes. CNV in SW620 cells was assessed by qPCR on the HSATII–chr16-1 locus (*C*), HSATII–chr16-2 locus (*D*), and chromosome 16q arm (*E*). Cycle threshold values for all samples were normalized against β-actin, and DNA CNV is expressed relative to SW620 cells before xenograft implants, which was set at 1 (T2, T6, T10 = 1 wk of culture after the second, sixth, and 10th serial transplants, respectively). Error bars represent SD ($n = 3$).

retrovirus (HERV) family members, and human telomerase reverse transcriptase (hTERT). Definitive identification of the cellular RTs responsible for HSATII RT is complicated by the diversity of the retrovirally encoded enzyme families and the limited reagents available for their analysis. Attempts to reduce HERV and LINE-1 levels in our cell line models with previously published siRNAs (23) were unsuccessful. Because tools to evaluate hTERT were readily available, we undertook RNA immunoprecipitation analysis of endogenous hTERT from nuclear extracts of SW620 and HCT116 cell lines, followed by RT-qPCR for HSATII species. Significant enrichment of HSATII RNA with hTERT immunoprecipitates was evident, relative to control IgG (*SI Appendix*, Fig. S4 *A* and *B*). Importantly, this enrichment was only observed when the cancer cells were grown as xenografts, but not when they were cultured under standard 2D in vitro conditions. As a control, the coprecipitation with hTERT of telomerase RNA component (TERC), its primary RNA template for telomere elongation, was unaffected by growth conditions. Negligible coprecipitation with hTERT was observed for an unrelated noncoding RNA target. Furthermore, TERT knockdown using three independent siRNAs demonstrated significant reduction, although not complete depletion, of HSATII rdDNA species (*SI Appendix*, Fig. S4 *C* and *D*). Thus, hTERT may mediate HSATII RT, although the contribution of other known cellular RTs cannot be excluded.

Progressive Expansion of Pericentromeric Loci Through Stable Reintegration of HSATII DNA Sequences. HSATII-derived rdDNA fragments within the cell nuclear fraction (*SI Appendix*, Fig. S2*G*) may either give rise to extrachromosomal elements or be integrated at chromosomal loci, leading to stable expansion of HSATII genomic sequences. By analogy, RT of LINE-1 transcripts, followed by retrotransposition at chromosomal loci, has been described in epithelial cancers, including colon carcinoma (24). To address this possibility, we first analyzed the dynamics of global HSATII RNA- and DNA-level changes using single molecule sequencing in SW620 colon cancer cells transitioned from 2D in vitro culture conditions to growth as mouse xenografts, and vice versa. As expected, the number of HSATII RNA reads was minimal when cells were cultured in 2D conditions, induced 360-fold as the cells gave rise to xenografts in mice, and then promptly down-regulated as xenograft-derived tumor cells were returned to in vitro 2D cultures (Fig. 4*A*, blue solid line). Remarkably, total cellular HSATII DNA copy number, which was already abundant at baseline, increased by 25-fold as 2D-cultured cells were transitioned to xenografts (taking into account the multiple-length variants of the HSATII tandem repeat unit). The amplified HSATII DNA sequences remained stably expanded despite the down-regulation of HSATII RNA transcripts when cells were reestablished under 2D culture conditions in vitro (Fig. 4*A*, red solid line). As a control, we analyzed gamma satellite II (GSATII), which is structurally similar to HSATII but whose expression is not deregulated in cancer (3). SW620 cells showed negligible GSATII changes, in either RNA or DNA content, as cells transitioned between 2D in vitro and xenograft culture conditions (Fig. 4*A*). Notably, the cancer-enriched human alpha satellite (ALR/Alpha) and simple satellite repeat (CATTC)n RNAs were also induced at high levels in xenografts with conjugate DNA copy number gains, suggesting a common mechanism of RT-mediated genomic expansion of particular repeat classes (*SI Appendix*, Table S1).

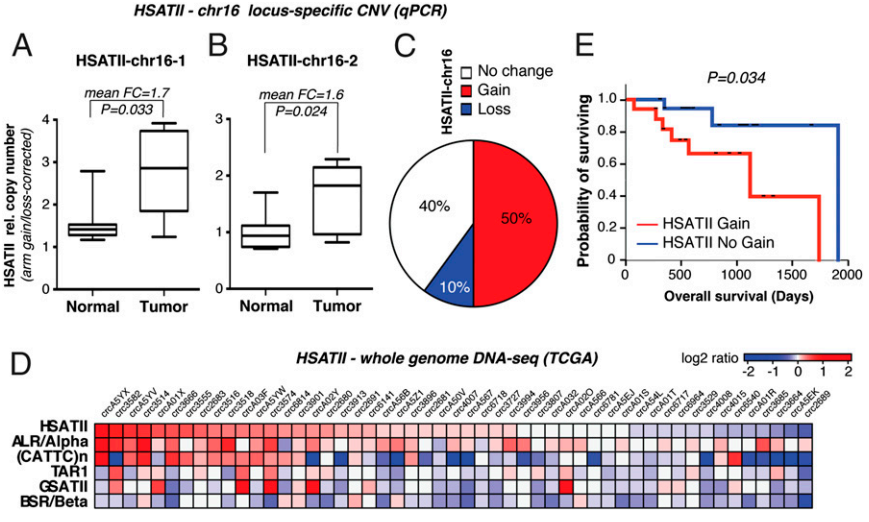


Fig. 5. Pericentromeric HSATII repeats expand both locally and genome-wide in primary human colon cancer samples. CNV analysis of HSATII–chr16-1 (*A*) and HSATII–chr16-2 (*B*) loci on the indicated paired colon specimens ($n = 10$) is shown. For each sample, values were normalized for β-actin DNA and corrected for chr16q arm changes. Probability was measured by the paired *t* test. FC, mean fold change. (*C*) Relative percentage of HSATII copy number changes in colon tumor/normal pairs according to combined HSATII–chr16-1 and HSATII–chr16-2 CNV analysis, including correction for chr16 arm gains/losses. (*D*) Heat map of whole-genome sequencing data on the indicated primary colon cancer specimens based on a $\log_2$ ratio cutoff of 0.1. (*E*) Kaplan–Meier curve of overall survival (days) of patients with primary colon cancers with HSATII CNV gain or no gain. $P = 0.034$ (log-rank *P* value).

To address the localization of the amplified HSATII DNA sequences, we performed HSATII DNA FISH analysis of 40 xenograft-derived metaphase spreads that did not reveal detectable extrachromosomal elements, and no hybridization signal was visible outside the five chromosomal loci known to harbor long arrays of pericentromeric HSATII (Fig. 4*B*). We could not discern increased fluorescence intensity or size of hybridization signal to demonstrate local expansion due to the limitations of this assay. However, consistent with the FISH data, alignment of HSATII gDNA reads obtained by single molecule sequencing from SW620 xenografts showed that the additional HSATII sequences were distributed among the various endogenous preexisting HSATII pericentromeric loci (*SI Appendix*, Fig. S5 *A* and *B*).

To model HSATII DNA copy gain over time and as a function of tumor progression, we serially transplanted SW620 cells as xenografts over multiple generations of mice. Progressive amplification of HSATII gDNA was evident over 10 successive rounds of in vivo tumor initiation, as measured in cultures derived from xenografted cells. This amplification was assessed using an ∼170-bp qPCR-based copy number variation (CNV) assay at the two highest density HSATII pericentromeric regions on chromosome 16q (HSATII–chr16-1 and HSATII–chr16-2; Fig. 4 *C* and *D* and *SI Appendix*, Fig. S5 *C* and *D*). An adjacent chromosomal region showed no xenograft transplantation-associated copy number changes, ruling out nonspecific gains in the 16q chromosomal arm or in ploidy (Fig. 4*E* and *SI Appendix*, Fig. S5*E*). We did not observe a single chromosome locus duplication event during tumor progression leading to a discrete increase in HSATII gDNA. Instead, the pericentromeric genomic loci demonstrated a gradual increase in HSATII gene copy number over time, all within preexisting satellite domains. Such a time line would be consistent with multiple rdDNA-mediated reintegration events.

Common Pericentromeric Expansion of HSATII Repeats in Human Colorectal Cancers. To determine whether HSATII copy number gains occur in primary human colon cancer, we analyzed CNV in 10 pairs of primary tumors and their matched adjacent normal tissue, focusing again on the chromosome 16q (HSATII–chr16-1 and HSATII–chr16-2) loci. After correcting for chr16q arm loss or gain, significantly increased HSATII copy number was evident at either or both of the two independent HSATII loci tested in five of 10 (50%) colon cancers (Fig. 5 *A*–*C* and *SI Appendix*, Fig. S6*A*). Among other cancers similarly analyzed, HSATII gene copy gain was evident in five of 13 (38%) kidney cancers (*SI Appendix*, Fig. S6*B*).

To extend our study of HSATII gene copy changes at specific loci, we performed a genome-wide survey of all such satellite repeats using a novel satellite CNV algorithm to undertake computational analyses of whole-genome sequencing from The Cancer Genome Atlas Project (TCGA). After correction of these data for large genomic alterations, comparable in size to HSATII stretches, we found that in fully annotated genomic sequences of 51 colorectal cancers (Dataset S1), 23 (45%) had statistically significant genomic gain of HSATII compared with their matched normal germ line (Fig. 5*D*). We extended this analysis to include additional satellite repeats, which revealed higher copy number gains in particular in the satellites whose expression was enriched in cancers [ALR/Alpha, HSATII, and (CATTC)n], but not in GSATII and other repeats whose expression is not deregulated in cancer cells (Fig. 5*D* and *SI Appendix*, Fig. S6*C*). HSATII copy gain co-occurred more frequently with ALR/Alpha and (CATTC)n, indicating a common mechanism for repeat expansion in some groups of repeats that is not shared with others (*SI Appendix*, Fig. S6*C*). In addition, alignments indicate that repeat expansions occur in the same locations consistent with our xenograft models. Of the 51 TCGA samples, 46 had annotated overall survival data that we analyzed to compare tumors with HSATII gain vs. no gain. Notably, Kaplan–Meier analysis demonstrated a significant reduction in overall survival in the HSATII gain vs. no-gain tumors (Fig. 5*E*; median overall survival: 1,096 vs. 1,881 d; log-rank *P* value = 0.034). Taken together, our data show that gene copy gains at HSATII-encoding pericentromeric repeats and other cancer-enriched satellites occur at preexisting repeat arrays and are a common and negative prognostic feature of colorectal cancers.

The mechanism underlying HSATII satellite repeat expansion in cancer remains to be defined, but it may involve an RT/reintegration phenomenon analogous to the phenomenon described for other major repetitive elements, such as LINE-1 (24) and telomeres (25, 26). Although the RT-directed expansion of pericentromeric sequences has not been described in human cells, there is ample precedent for retroelement-mediated integration of such repeats in plants (27, 28). In mammalian cells, integration of nonretroelement sequences within centromeres has not been reported, with the exception of the marsupial

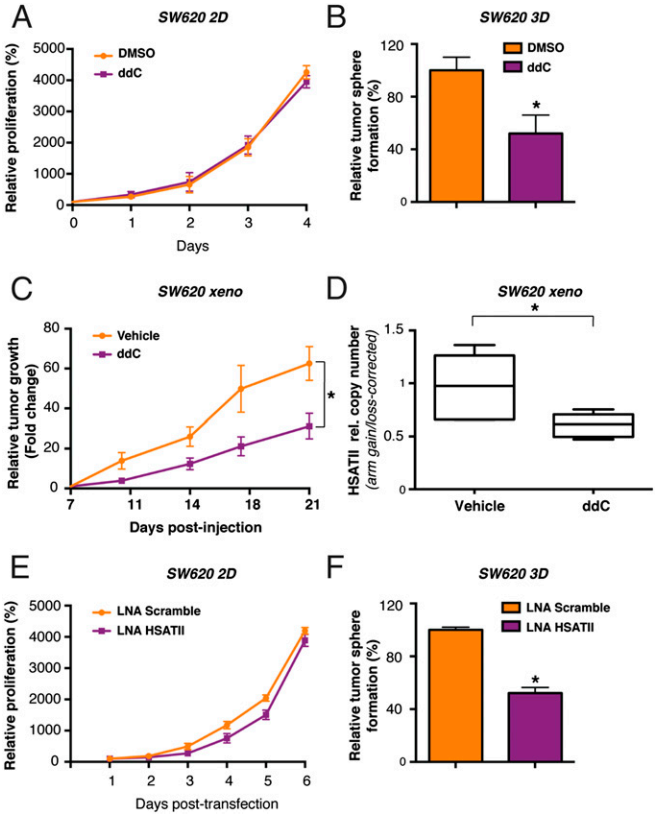


Fig. 6. RT blockage, as well as LNA-mediated inhibition of HSATII transcripts, affects tumor sphere growth, impairs tumorigenesis, and prevents pericentromeric copy number gains in vivo. (*A*) Proliferation assay on DMSO- and ddC-treated SW620 cells. Values (%) were normalized against the signal derived from viable cells on the day of seeding, which was set at 100%. (*B*) Tumor sphere-forming ability of SW620 cells was tested upon culture in the presence of ddC for 15 d. Values (%) were normalized against the amount of spheres in the DMSO control, which was set at 100%. **P* < 0.05 (*t* test). (*C*) In vivo tumor growth of SW620 cell xenografts. Mice were treated daily by i.p. injection of 25 mg/kg ddC or vehicle alone, starting 1 wk after tumor cell injection. Tumor size at this stage was set at 1 to calculate relative size fold change over time. Error bars represent SEM ($n = 6$). **P* < 0.05 (*t* test). (*D*) CNV analysis of HSATII by qPCR on the chr16-1 locus in tumors recovered from untreated (Vehicle) or ddC-treated mice ($n = 6$). Values were normalized for β-actin and expressed as HSATII/chr16q arm ratios. **P* < 0.05 (*t* test). (*E*) Proliferation assay on SW620 cells upon transfection with an HSATII-specific or Scramble LNA. Values (%) were normalized against the signal derived from viable cells 1 d after transfection, which was set at 100%. (*F*) Tumor sphere-forming ability of SW620 cells was tested upon transfection with an HSATII-specific or Scramble LNA. Values (%) were normalized against the amount of spheres in the control, which was set at 100%. **P* < 0.05 (*t* test). In *A*, *B*, *E*, and *F*, error bars represent SD ($n = 3$).

tammar wallaby, whose exceptionally short centromeres harbor signatures of retroviral insertions alongside domains of satellite-rich sequences (29). Thus, by analogy with retroviral elements, LINE-1, and telomeres, pericentromeric repeats may expand through the activity of endogenous RT enzymes. However, we cannot unequivocally discriminate between candidate RTs or exclude other mechanisms, including replication-induced errors or epigenetic modifier-induced site-specific copy gain (30–32).

Suppression of HSATII RT Inhibits Tumor Growth and Impairs Copy Number Gains. The critical role of pericentromeric repeats in preserving chromosomal integrity is well established (1). However, the unexpected RT-associated mechanism by which pericentromeres are expanded in cancer cells raises the possibility that its disruption may affect tumor growth. Given the inhibition of HSATII rdDNA formation by treatment of cells with the nucleoside analog ddC, we first tested the effect of this NRTI on cell proliferation. Treatment of SW620 cells with ddC inhibited their proliferation under 3D conditions but had minimal effect in 2D culture (Fig. 6 *A* and *B*). Mouse tumor xenografts generated from SW620 cells also showed sensitivity to ddC, with a 50% reduction in tumor diameter at 21 d ($P = 0.03$; Fig. 6*C*). The antiproliferative effect of ddC was accompanied by a reduction of HSATII copy gain in tumor xenografts (Fig. 6*D*). In a second colon cancer cell line, HCT116, the growth of tumor xenografts was suppressed using a combination of two RT inhibitors, ddC and 2′,3′-didehydro-2′,3′-dideoxythymidine (d4T), but not using ddC alone, and this effect was again associated with a reduced HSATII copy number gain (*SI Appendix*, Fig. S7 *A–C*). To test an alternative strategy to target HSATII, we synthesized a locked nucleic acid (LNA) oligonucleotide complementary to the HSATII sequence. Treatment of SW620 cells with this HSATII-directed LNA had no effect when the cells were grown in two dimensions, but it had a strong inhibitory effect on tumor sphere formation (Fig. 6 *E* and *F*). This effect was associated with accumulation of cells in the G0/G1 phase of the cell cycle (*SI Appendix*, Fig. S7*D*). Taken together, these results raise the intriguing possibility that targeting HSATII transcripts or suppressing cellular reverse transcriptase activity may selectively suppress proliferation of cancer cells under anchorage-independent conditions (*SI Appendix*, Fig. S7*E*). Beyond their potential contribution to the viability of proliferating cancer cells, HSATII transcripts may also play a role in shaping tumor–host interactions, as shown in the accompanying paper (33). Thus, the disruption of HSATII RT may have direct effects on cancer cells, as well as modulating the immune response against tumor cells. Given the very high frequency of HSATII deregulation in epithelial cancers (3), such a therapeutic vulnerability might have broad significance.

Materials and Methods

Human normal and tumor tissues were deidentified and discarded excess tissue obtained from the Massachusetts General Hospital (MGH) according to an MGH Institutional Review Board (IRB)-approved protocol (2013P001854). The IRB determined consent was not needed for this study. Total RNA from normal human pancreas was purchased from Clontech. The results here are based, in part, upon data generated by the TCGA Research Network (cancergenome.nih.gov/). TCGA data are available from the TCGA portal. Helicos sequencing data were only analyzed for satellite expression, which is shown in *SI Appendix*. Given that it is not standard expression, we did not deposit the sequence into National Center for Biotechnology Information Gene Expression Omnibus. Further experimental details are provided in *SI Appendix*.

ACKNOWLEDGMENTS. We thank M. Miri from the Massachusetts General Hospital (MGH) Tissue Repository for providing pathological specimens. We also thank N. J. Dyson and M. R. Motamedi for critically revising the manuscript, as well as R. Taulli and all laboratory members for helpful discussions. This work was supported by the Howard Hughes Medical Institute (D.A.H.), the National Foundation for Cancer Research (D.A.H.), National Cancer Institute (NCI) Grant R01CA129933 (to D.A.H. and F.B.), the Burroughs Wellcome Trust (D.T.T.), NIH Grant K12CA087723-11A1 (to D.T.T.), Department of Defense Grant W81XWH-13-1-0237 (to D.T.T.), the Warshaw Institute for Pancreatic Cancer Research (D.T.T.), the Verville Family Pancreatic Cancer Research Fund (D.T.T.), the Eleanor and Miles Shore Fellowship (to E.L.), the William Randolph Hearst Fund (E.L.), Susan G. Komen for the Cure Grant KG09042 (to S.M.), and the NCI Federal Share Program and Income (S.M.).

1. Bierhoff H, Postepska-Igielska A, Grummt I (2013) Noisy silence: Non-coding RNA and heterochromatin formation at repetitive elements. Epigenetics 9(1):53–61.
2. Eymery A, Callanan M, Vourc'h C (2009) The secret message of heterochromatin: New insights into the mechanisms and function of centromeric and pericentric repeat sequence transcription. *Int J Dev Biol* 53(2-3):259–268.
3. Ting DT, et al. (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331(6017):593–596.
4. Eymery A, et al. (2009) A transcriptomic analysis of human centromeric and pericentric sequences in normal and tumor cells. *Nucleic Acids Res* 37(19):6340–6354.
5. Zhu Q, et al. (2011) BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* 477(7363):179–184.
6. Jeanpierre M (1994) Human satellites 2 and 3. *Ann Genet* 37(4):163–171.
7. Ehrlich M (2009) DNA hypomethylation in cancer cells. *Epigenomics* 1(2):239–259.
8. Tilman G, et al. (2012) Cancer-linked satellite 2 DNA hypomethylation does not regulate Sat2 non-coding RNA expression and is initiated by heat shock pathway activation. *Epigenetics* 7(8):903–913.
9. Lipson D, et al. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 27(7):652–658.
10. Bouzinba-Segard H, Guais A, Francastel C (2006) Accumulation of small murine minor satellite transcripts leads to impaired centromeric architecture and function. *Proc Natl Acad Sci USA* 103(23):8709–8714.
11. Rizzi N, et al. (2004) Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. *Mol Biol Cell* 15(2):543–551.
12. Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J (2007) Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* 318(5851):798–801.
13. Valgardsdottir R, et al. (2008) Transcription of Satellite III non-coding RNAs is a general stress response in human cells. *Nucleic Acids Res* 36(2):423–434.
14. Leonova KI, et al. (2013) p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. *Proc Natl Acad Sci USA* 110(1):E89–E98.
15. Levin HL, Moran JV (2011) Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* 12(9):615–627.
16. Boguslawski SJ, et al. (1986) Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. *J Immunol Methods* 89(1): 123–130.
17. Huertas P, Aguilera A (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol Cell* 12(3):711–721.
18. Hu Z, Zhang A, Storz G, Gottesman S, Leppla SH (2006) An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res* 34(7):e52.
19. Rigby RE, et al. (2014) RNA:DNA hybrids are a novel molecular pattern sensed by TLR9. *EMBO J* 33(6):542–558.
20. Kukhanova M, et al. (1995) L- and D-enantiomers of 2′,3′-dideoxycytidine 5′-triphosphate analogs as substrates for human DNA polymerases. Implications for the mechanism of toxicity. *J Biol Chem* 270(39):23055–23059.
21. Louat T, et al. (2001) Antitumor activity of 2′,3′-dideoxycytidine nucleotide analog against tumors up-regulating DNA polymerase beta. *Mol Pharmacol* 60(3):553–558.
22. Dai L, Huang Q, Boeke JD (2011) Effect of reverse transcriptase inhibitors on LINE-1 and Ty1 reverse transcriptase activities and on LINE-1 retrotransposition. *BMC Biochem* 12:18.
23. Oricchio E, et al. (2007) Distinct roles for LINE-1 and HERV-K retroelements in cell proliferation, differentiation and tumor progression. *Oncogene* 26(29):4226–4233.
24. Lee E, et al.; Cancer Genome Atlas Research Network (2012) Landscape of somatic retrotransposition in human cancers. *Science* 337(6097):967–971.
25. Hastie ND, et al. (1990) Telomere reduction in human colorectal carcinoma and with ageing. *Nature* 346(6287):866–868.
26. Counter CM, et al. (1992) Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *EMBO J* 11(5): 1921–1929.
27. Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromeres. *Trends Plant Sci* 8(12):570–575.
28. Neumann P, et al. (2011) Plant centromeric retrotransposons: A structural and cytogenetic perspective. *Mob DNA* 2(1):4.
29. Carone DM, et al. (2009) A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* 118(1):113–125.
30. Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* 191(4227):528–535.
31. Cohen Z, Bacharach E, Lavi S (2006) Mouse major satellite DNA is prone to eccDNA formation via DNA Ligase IV-dependent pathway. *Oncogene* 25(33):4515–4524.
32. Black JC, et al. (2013) KDM4A lysine demethylase induces site-specific copy gain and rereplication of regions amplified in tumors. *Cell* 154(3):541–555.
33. Tanne A, et al. (2015) Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells. *Proc Natl Acad Sci USA*, 10.1073/pnas.1517584112.

Distinguishing the immunostimulatory properties of noncoding RNAs expressed in cancer cells

Antoine Tanne[a], Luciana R. Muniz[a], Anna Puzio-Kuter[b], Katerina I. Leonova[c], Andrei V. Gudkov[c], David T. Ting[d], Rémi Monasson[e], Simona Cocco[f], Arnold J. Levine[b,g,1], Nina Bhardwaj[a,2], and Benjamin D. Greenbaum[a,g,h,1,2]

[a]Tisch Cancer Institute, Department of Medicine, Hematology, and Medical Oncology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [b]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903; [c]Roswell Park Cancer Institute, Buffalo, NY 14263; [d]Massachusetts General Hospital, Charlestown, MA 02129; [e]Laboratoire de Physique Théorique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [f]Laboratoire de Physique Statistique, CNRS and Ecole Normale Supérieure, 75005 Paris, France; [g]The Simons Center for Systems Biology, School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and [h]Department of Pathology, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Contributed by Arnold J. Levine, September 10, 2015 (sent for review April 27, 2015; reviewed by Chakraborty Arup and Curtis G. Callan Jr.)

Recent studies have demonstrated abundant transcription of a set of noncoding RNAs (ncRNAs) preferentially within tumors as opposed to normal tissue. Using an approach from statistical physics, we quantify global transcriptome-wide motif use for the first time, to our knowledge, in human and murine ncRNAs, determining that most have motif use consistent with the coding genome. However, an outlier subset of tumor-associated ncRNAs, typically of recent evolutionary origin, has motif use that is often indicative of pathogen-associated RNA. For instance, we show that the tumor-associated human repeat human satellite repeat II (HSATII) is enriched in motifs containing CpG dinucleotides in AU-rich contexts that most of the human genome and human adapted viruses have evolved to avoid. We demonstrate that a key subset of these ncRNAs functions as immunostimulatory "self-agonists" and directly activates cells of the mononuclear phagocytic system to produce proinflammatory cytokines. These ncRNAs arise from endogenous repetitive elements that are normally silenced, yet are often very highly expressed in cancers. We propose that the innate response in tumors may partially originate from direct interaction of immunogenic ncRNAs expressed in cancer cells with innate pattern recognition receptors, and thereby assign a previously unidentified danger-associated function to a set of dark matter repetitive elements. These findings potentially reconcile several observations concerning the role of ncRNA expression in cancers and their relationship to the tumor microenvironment.

noncoding RNA | genome evolution | cancer immunology

The recent development of total RNA sequencing has allowed a better appreciation of the complexity and breadth of the entire transcriptome (1–4). Analysis by the Encyclopedia of DNA Elements (ENCODE) consortium unexpectedly showed that far more of the mammalian genome than previously appreciated is transcribed into noncoding RNA (ncRNA). Several short ncRNAs have conserved metabolic and regulatory functions, and some antiviral properties have been assigned to novel ncRNA classes, such as eukaryotic siRNA, piRNA (PIWI-interacting) RNA, and prokaryotic CRISPR (clustered regularly interspaced short palindromic repeats) RNA (5). In eukaryotes, long noncoding RNA (lncRNA), such as long-intergenic ncRNA, has been associated with transcriptional, posttranscriptional, and epigenetic regulation (6, 7).

It is now evident that germ-line and cancer cells can have atypical ncRNA transcription, including repetitive elements from regions usually silenced in steady state (8, 9). In eukaryotes, transcription of endogenous retroviruses and mobile elements is mostly repressed epigenetically through processes such as histone modification and DNA methylation, preventing disruptive or deregulatory effects due to integration into coding regions. In mammals, DNA methylation targets the cytosine in CpG motifs to form 5-methylcytosine contributing to down-regulation of transcription for methylated sequences (10). Epigenetic regulation is strongly associated with the developmental process, whereas its deregulation, such as by disruption of DNA methylation, can be associated with dedifferentiation and carcinogenic processes (11, 12).

In cancers, such as those cancers driven by p53 mutations and epigenetic alterations, ncRNA associated with repetitive elements can be induced (8, 9). In a study of mouse and human epithelial malignancies by Ting et al. (9), several repetitive elements emanating from genomic dark matter and often repressed in steady-state conditions, particularly in pericentromeric repeats, such as GSAT (major satellite) in mouse and human satellite repeat II (HSATII) in humans, were only transcribed in cancer cells. Leonova et al. (8) demonstrated a strong induction of repetitive elements from the mouse genome (particularly GSAT, B1, and B2), along with several other ncRNAs, in cells bearing p53 oncogenic mutations and exposed to epigenome-altering demethylating agents. Anomalous expression of the murine repetitive element GSAT was shown to trigger transcription of the repeat-dependent activated IFN response, which can regulate apoptosis-related cell death. Similarly, when expressed, endogenous retroviral RNA can activate the innate immune response via several pathways (13). Altogether, these studies suggest that certain ncRNAs may also have attributes of immunostimulatory nucleic acid sequences.

We use a set of mathematical tools originally developed to analyze potentially immunostimulatory motif use in viral and host genome coding sequences. These methods were recently recast in the language of statistical physics and are extended here to analyze ncRNA motif use (14, 15). We analyze for the first time, to our knowledge, large-scale patterns of motif use in human and murine

Significance

Using an approach derived from statistical physics, we quantify transcriptome-wide motif usage in human and murine noncoding RNAs (ncRNAs), determining that most have motif usage consistent with the coding genome. However, an outlier subset of tumor-associated ncRNAs comprises repetitive elements whose motif usage patterns are more typically associated with the genomes of inflammatory pathogens. We demonstrate that a key subset of these elements directly activates the cellular innate immune response. We propose that the innate response in tumors partially originates from direct interaction of immunogenic ncRNAs preferentially expressed in cancer cells with innate pattern recognition receptors.

Author contributions: A.T., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. designed research; A.T., L.R.M., A.P.-K., R.M., S.C., and B.D.G. performed research; D.T.T., R.M., S.C., and B.D.G. contributed new reagents/analytic tools; A.T., L.R.M., A.P.-K., K.I.L., A.V.G., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. analyzed data; and A.T., D.T.T., R.M., S.C., A.J.L., N.B., and B.D.G. wrote the paper.

Reviewers: C.A., Massachusetts Institute of Technology; and C.G.C., Princeton University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. Email: alevine@ias.edu or benjamin.greenbaum@mssm.edu.

[2]N.B. and B.D.G. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1517584112/-/DCSupplemental.

transcriptomes, which we use to find anomalies in ncRNA expressed in cancer transcriptomes (5, 16). As a result, we are able to characterize features of ncRNA overexpressed in cancerous cells relative to normal cells (8, 9, 17). Our analysis includes several large datasets of functionally characterized ncRNA, in addition to pseudogenes and repetitive elements, such as satellite DNA, endogenous retroviruses, and long and short interspersed elements. We demonstrate many ncRNAs preferentially expressed in cancerous cells display anomalous motif use patterns compared with the vast majority of ncRNAs whose patterns of motif use we show to be consistent with those patterns of motif use in coding regions. Based on their unusual pattern of motif use and differential expression in cancerous vs. normal cells, we predicted that HSATII and GSAT incorporate immunostimulatory motifs in humans and mice, respectively. Remarkably, we validate our prediction demonstrating that both directly stimulate antigen-presenting cells and accordingly label them immunostimulatory ncRNAs (i-ncRNAs).

Results

General Motif Use Patterns in lncRNAs. Using the GENCODE database of lncRNA transcripts from humans and mice (versions 19 and 2 for humans and mice, respectively) we calculated the strength of statistical bias (referred to as a force) on sequence motif use for all contained lncRNAs as described in *Materials and Methods*. GENCODE lncRNA established a baseline of sequence motif use expressed in a broad array of cells and tissues so that we could compare these patterns of motif use with those patterns of motif use of ncRNAs expressed in certain cancers. For each sequence, we calculate the force on all two- and three-nucleotide motifs and use Eq. **5** in *Materials and Methods* to calculate the probability of observing a sequence with that number of motifs. The number of sequences in GENCODE for which a given dinucleotide is aberrantly expressed is illustrated in Fig. 1*A*. CpG dinucleotides are vastly underrepresented, as indicated by their negative forces in *SI Appendix*, Table S1. UpA dinucleotides are often underrepresented, although to a lesser extent. As in our previous work, these patterns cannot be explained by nucleotide frequencies, such as guanine-cytosine (GC) content, which are accounted and normalized for in our method.

These dinucleotide motif use patterns are similar in human and mouse genomes across the wide array of cells and cell lines contained in GENCODE (2, 3). Strikingly, avoidance of the CpG and UpA dinucleotide motifs in this dataset is stronger than in coding regions (*SI Appendix*, Fig. S1). One can conclude that the patterns previously observed in virus and host coding genes are not due to effects from coding regions, such as codon use patterns (18–20). Rather, such constraints in coding regions likely weaken the strength of a statistical bias that comes from the same underlying mechanisms. This pattern suggests selective restrictions on dinucleotide frequencies observed in ncRNAs preserving a function or avoiding a detrimental consequence, such as a chronic autoinflammatory response that could result from presenting danger-associated molecular patterns (DAMPs). Adaptation of dinucleotide motif use in these elements over time is analogous to the viral mimicry of host patterns of sequence motif use (14, 21). When an avian influenza virus enters the human population, one can observe adaptation to analogous patterns emerging over time (14, 15, 22, 23). In that case, mutation rates in influenza are very high, so one can follow these evolutionary adaptations over far shorter time periods.

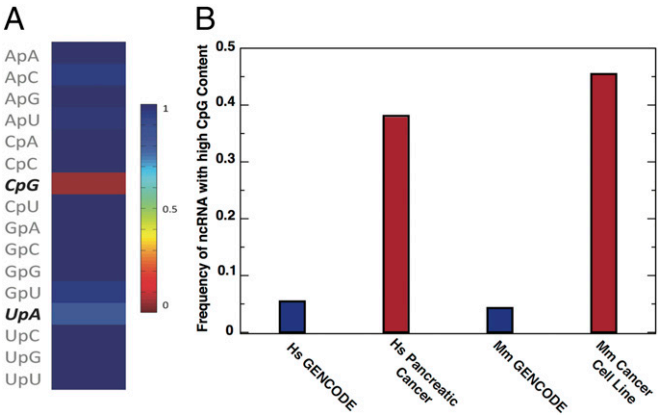


Fig. 1. ncRNAs expressed in cancer differ from general lncRNA motif use patterns. (*A*) Fraction of GENCODE human lncRNA sequences where a motif occurs the expected number of times as defined by corresponding to a probability greater than 0.05 (Eq. **5**). (*B*) Fraction of GENCODE lncRNA sequences in humans (Hs) and mice (Mm) where CpG motifs occur the expected number of times compared with the CpG motifs expressed in human cancerous cells and mouse cancer cell lines.

Trinucleotide motifs with significant forces are listed in the *SI Appendix*, Table S1, along with dinucleotide motifs. Trinucleotide motifs with significant forces acting on them are conserved between humans and mice, as was the case for dinucleotides, with the exception of UAC and UAG (which are significant in humans but less so in mice). Except for UAG (chain termination codons used in coding RNAs), whenever a trinucleotide motif is significantly enhanced or avoided in humans, its reverse complement is also significantly enhanced or avoided, suggesting avoidance of complementary motifs. The strongest forces suppress CpG and CpG-containing trinucleotides particularly when an A or U is next to the core CpG motif. These results are consistent with the avoidance of CpGs in AU contexts observed in influenza viruses replicating in humans (15, 22, 23). Given the apparent bias against CpG and UpA, we sought to determine if these motifs were linked. Pearson correlation between these forces across all GENCODE ncRNA in humans and mice showed no correlation between CpG and UpA biases (r = 0.0006; *SI Appendix*, Fig. S2). Therefore, the forces on CpG and UpA are likely independent. Moreover, every significant trimer across the GENCODE is correlated to CpG, UpA, or both. As a result, all significant trimers can be explained by their CpG or UpA motif use.

Cancer-Enriched Noncoding Repeat RNA May Have Anomalous Motif Use. Prior work revealed aberrant expression of ncRNA across a spectrum of mouse and human cancers (8, 9). These sequences were found in the Repbase database of human and murine repetitive elements and the Functional Annotation of Mouse (FANTOM) database of murine noncoding elements (currently NONCODE) (24, 25). We also found high induction of GSAT in a murine testicular teratoma and liposarcoma tumor model (8, 9) (*SI Appendix*, Fig. S3). Focusing on these cancer-expressed repeats, we found a surprisingly significant enrichment of anomalous motif use patterns compared with other ncRNAs. In the Repbase database, we tested whether the bias on dinucleotide and trinucleotide motifs observed in repetitive element sequences fell outside the distribution obtained from GENCODE lncRNA. Remarkably, we found hundreds of sequences falling outside of this distribution. Many have high use of CpG dinucleotides, including a set of endogenous viruses (*SI Appendix*, Table S2) recently implicated in the innate immune response in tumors (13). We conclude that although the portions of the noncoding regions typically expressed as lncRNAs have motif use patterns similar to RNA from coding regions, there are many genomic regions with atypical motif use that are not transcribed in normal cells or tissues.

We use the forces that quantify the strength of the statistical bias on the often underrepresented CpG and UpA dinucleotides to differentiate between ncRNAs found preferentially in cancerous cells and the total lncRNA referenced in GENCODE for humans and mice, because these two dinucleotides essentially account for all significant trinucleotide motifs in this set. We use the distribution of forces on CpG and UpA to define a null hypothesis, which we approximate by a Gaussian distribution (Fig. 2). Many ncRNAs from cancerous cells are clearly outside the distribution, often to a large extent. In particular, HSATII, the main ncRNA up-regulated in human pancreatic cancers, is far outside the human

distribution, and GSAT, the main murine ncRNA implicated in murine tumoral cell lines, is well outside the mouse distribution. Within our null hypothesis, the P values for all ncRNAs considered here are less than 10^{-61} for human pancreatic cancer data and less than 10^{-2} for murine cell line data.

Many of the ncRNAs from the studies of Leonova et al. (8) and Ting, et al. (9) are outliers of at least three SDs with respect to at least one of the significant motifs implicated in the previous section, accounting for a median of 70.86% of the modulated Repbase RNA expression induced in pancreatic cancer, along with even higher percentages (73.95% and 85.74%, respectively) in the smaller sets of prostate and lung cancers. HSATII is the most differentially expressed (by a considerable margin) in the pancreatic cancer data, and HSATII and BSR are the highest in prostate and lung cancer data. In p53 KO murine cell lines treated with demethylation agents, around 68 ncRNAs are significantly modulated (8). Among those ncRNAs, 79.03% of the total expression comes from outliers as defined above, with the vast majority coming from GSAT and B2. Overall, we observed that repetitive sequences containing unusual motif use had varying degrees of conservation. However, the subset preferentially expressed in cancerous cells and tissues is encoded by sequences of more recent evolutionary origin. HSATII and GSAT are only conserved back to primates and mice, respectively, and 21 of the 22 ncRNAs from the study of Ting et al. (9) are conserved in humans and primates but extend no further back in evolution. Any function is likely to be species-specific.

ncRNAs with Unusual Motif Use Highly Expressed in Cancers Are Immunostimulatory. Our analysis highlights that many ncRNAs up-regulated in cancer display abnormal nucleotide motif use that we had previously related to immunogenic properties in viruses. The innate immune system contains several effector cells that react to immunogenic nucleic acids, such as exogenous viral and bacterial nucleic acids, as well as endogenous nucleic acids that can be released upon cell death (6). Among those effectors, the mononuclear phagocytic system [macrophages, monocytes, and dendritic cells (DCs)] contains key regulators of innate immune activation and adaptive immunity (26–28). DCs efficiently sense and sample their environment to integrate information and mount a proper response, which may be tolerogenic or immunogenic. To test whether ncRNA with highly unusual motif use could be recognized as a DAMP by some nucleic acid-sensing pattern recognition receptors (PRRs), we studied the effect of human HSATII and murine GSAT following transfection in human monocyte-derived DCs (moDCs) and murine bone marrow-derived macrophages. Liposomal transfection was required for stimulation, whereas naked RNA had no effect, implying recognition is consistent with activation via an endosomal or intracellular sensor (*SI Appendix*, Fig. S4). The general sets of recognition pathways tested are indicated in the *SI Appendix*, Fig. S5.

We generated different ncRNAs by in vitro transcription using minigenes coding for the two main candidate outliers computationally predicted to have immunogenic motif use (HSATII and GSAT). As controls, we derived RNA from minigenes encoding scrambled (sc) versions with the same nucleotide content but having normal motif use (labeled HSATII-sc and GSAT-sc) and repetitive elements of comparable length but having normal motif use patterns (RMER16A3 and UCON38), as described in *SI Appendix*. In human moDCs, liposomal transfection of HSATII induced significant production of IL-6, IL-12, and TNF-alpha relative to both endogenous controls and their scrambled versions (Fig. 3*A*). A similar profile of cytokines was elicited by moDCs in response to selected Toll-like receptor (TLR) agonists (*SI Appendix*, Fig. S6*A*). The candidate murine immunogenic ncRNA, GSAT, had less pronounced immunogenic properties but still induced IL-12 (Fig. 3*A*). Upon liposomal transfection of the same ncRNA into immortalized murine bone marrow-derived macrophages (imBMs), the immunogenic properties of HSATII were strongly attenuated, whereas the murine GSAT induced high levels of TNF-alpha (Fig. 3*B*) and monocyte chemotactic protein 1 (MCP-1), but not IFN-gamma, IL-6, or IL-12. The imBM almost exclusively regulates TNF-alpha in response to PRR agonists (*SI Appendix*, Fig. S6*B*).

HSATII and GSAT ncRNA induced IL-12 in human moDCs similar to the TLR3 ligand poly-IC (a synthetic dsRNA mimic; *SI Appendix*, Fig. S5). The absence of an effect by ncRNA with normal motif use [i.e., the scrambled forms (Fig. 3 *A* and *B*)] suggests specific sequence patterns within the RNA, such as CpG and UpA motifs, regulate immunostimulatory activity. Such motif use could

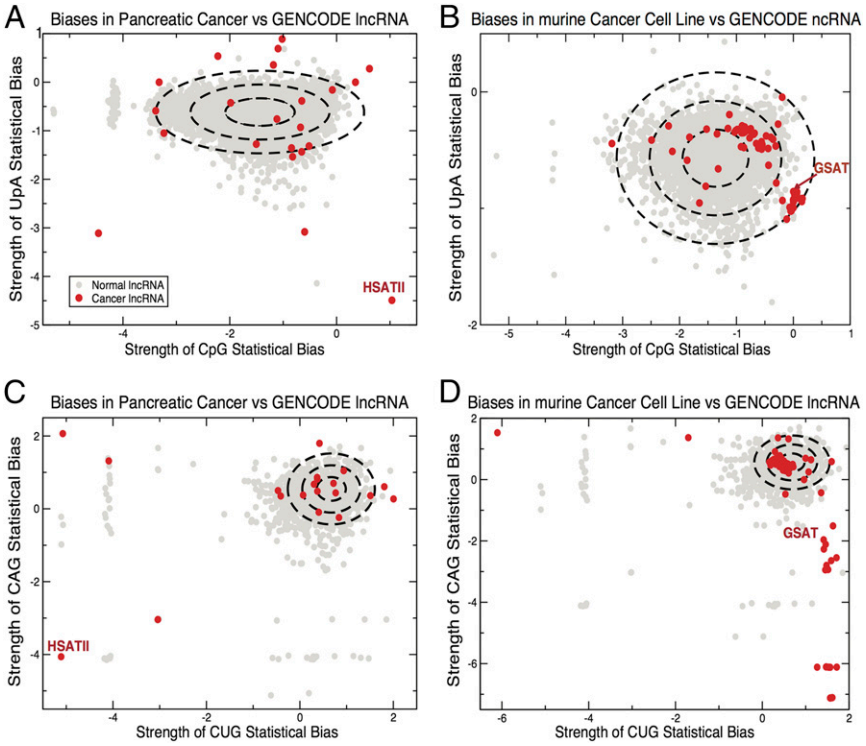


Fig. 2. ncRNA from cancer cells contains outliers from normal motif use. Distribution of UpA and CpG bias in lncRNA taken from human tumors (*A*) and murine cell lines (*B*) (indicated in red) plotted against lncRNA from GENCODE (indicated in gray). Each ellipse indicates 1 SD from the mean value in the GENCODE dataset. The forces on CAG and CUG are also shown for human tumors (*C*) and murine cell lines (*D*).

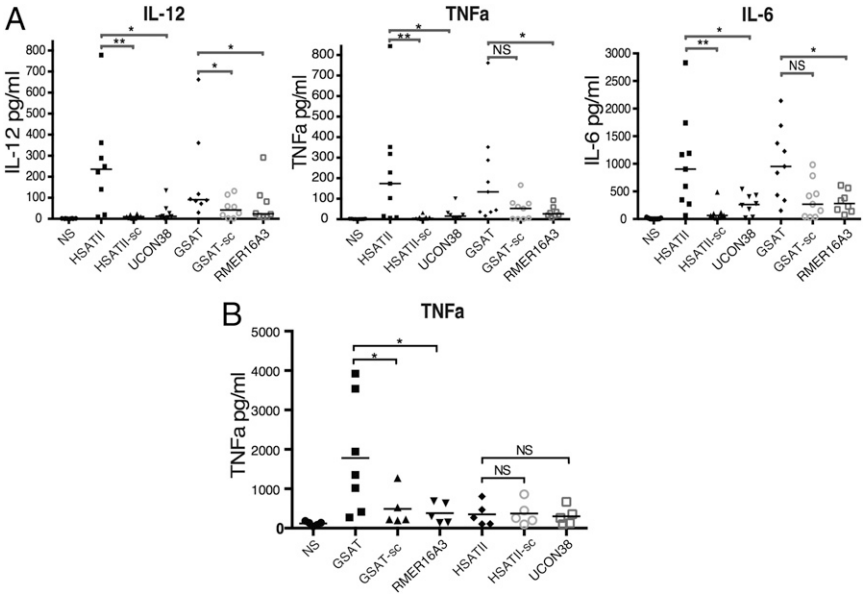


Fig. 3. i-ncRNA stimulates human moDC cytokine production. Quantification of inflammatory cytokine production in human moDCs (*A*) and murine imBMs (*B*) upon liposomal transfection of human i-ncRNA (HSATII) and murine i-ncRNA (GSAT) vs. their scrambled and endogenous controls. Each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median. The significance of i-ncRNA stimulation is analyzed by the nonparametric Mann–Whitney test to compare their effect vs. their scrambled and endogenous controls. NS, not significant. $*P < 0.05$; $**P < 0.01$.

also influence secondary conformations that may contribute to immunogenic properties, although we checked that the scrambled sequences did not lower the RNA minimum folding energy. Based upon these observations, we refer to HSATII and GSAT as immunogenic ncRNA or i-ncRNA. Interestingly, our study corroborates previous findings by Leonova et al. (8) that ncRNA, such as GSAT, can induce an innate response, although in those studies, the type I IFN pathway was also activated. Our initial investigations into this pathway were inconclusive (*SI Appendix*, Fig. S6*C*).

Dissection of the Immunostimulatory Properties of i-ncRNA. Pathogen-associated molecular patterns and DAMPs activate innate immune cells through PRRs. To characterize better the mechanisms involved in sensing i-ncRNA, we studied the immunomodulatory properties of HSATII and GSAT on a panel of imBMs that lack specific PRRs or effector molecules in their downstream signaling pathways (*SI Appendix*, Fig. S5). Whereas GSAT induced a TNF-alpha response, HSATII did not induce differential cytokine expression in these immortalized cells, indicating that there is either a species-specific effect, because the cells are murine, or a cell type-specific effect, because these cells are macrophages. This result is perhaps unsurprising, because different species and cell types express different PRRs, and HSATII and GSAT have different sequence compositions. Significantly, the absence of two key adaptor and regulatory proteins, MYD88 and UNC93B1: UNC93B3d (UNC93b), respectively, eliminated the differential response to GSAT in imBMs (Fig. 4).

MYD88 is a key cytosolic adaptor protein that is used by all TLRs except TLR3 to activate the transcription factor NF-κB. Similarly, the mutated form of UNC93b essentially eliminated inflammatory responses in imBMs. Although less well characterized than MYD88, this protein is known to interact with several endosomal TLRs (TLR3, TLR7, and TLR9) and has been implicated in TLR trafficking between the endoplasmic reticulum and endosomes, and their resultant maturation (29–31). We tested the requirement for TLR3, TLR7, and TLR9, which are known to recognize dsRNA, ssRNA, and CpG DNA, respectively (32–34) (*SI Appendix*, Fig. S7*A* and S8). None of these receptors were required for GSAT to activate TNF-alpha production from imBMs. Additional pathways investigated, including the stimulator of IFN genes (STING) and inflammasome pathways, are discussed in *SI Appendix* and did not contribute to i-ncRNA stimulatory activity. Altogether, our data are consistent with a requirement for i-ncRNA activation through signaling pathways that rely upon MYD88 and UNC93b. The precise receptor involved in initial recognition remains to be determined.

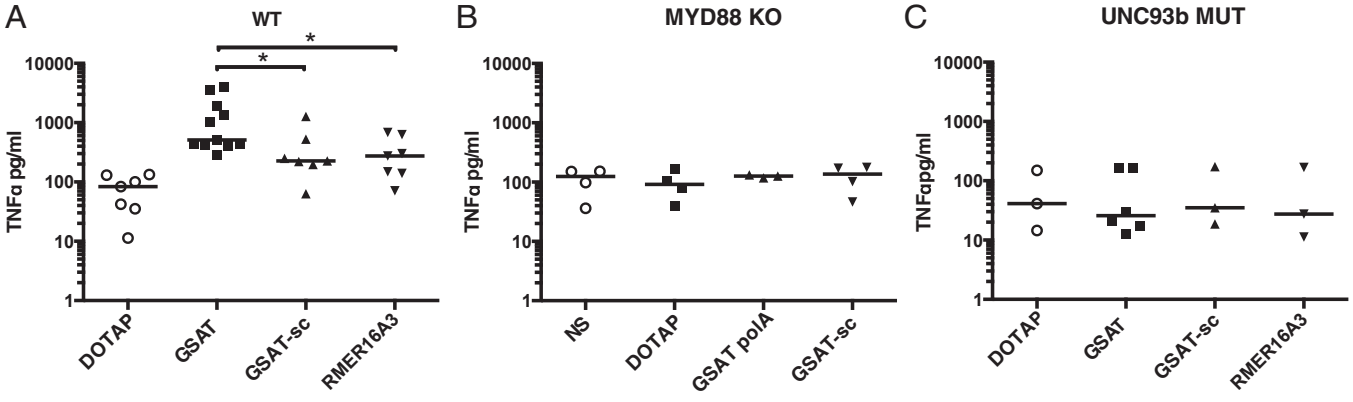


Fig. 4. MYD88 and UNC93b control GSAT i-ncRNA stimulation. Genetic screen of the innate immune pathway related to i-ncRNA function in murine imBMs. The imBM cells of different genotypes (WT, MYD88 KO, and UNC93b$^{3d/3d}$ MUT) have been stimulated by liposomal transfection (DOTAP liposomal transfection reagent) of the murine i-ncRNA (GSAT). TNF-alpha (TNF-a) production in the supernatant has been quantified, and each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median. $*P < 0.05$.

Discussion

There is a surprising similarity to be drawn between foreign viral nucleotide sequences and select ncRNAs silent in normal cells, yet transcribed in cancer cells, activating innate immunity (23, 29, 35–37). We determined that ncRNAs expressed predominantly in normal cells from humans and mice reflect patterns of nucleotide sequence motif avoidance, such as underrepresentation of CpG-containing sequences and reduced UpA, similar to protein-coding RNA. Such patterns often include a many-fold underrepresentation of CpG-containing sequences and reduced UpA motif use compared with expected levels. However, the genome also harbors repetitive elements, which often have abnormal use of CpG and UpA motifs compared with the use of CpG and UpA motifs observed in RNA expressed in normal cells and tissues. Sets of these ncRNAs, typically newer genome entries over evolutionary time scales, can be expressed at very high levels in cancerous cells and tumors. As a result, human and mouse elements expressed in cancer cells can have different sequences but can share high CpG content and are not generally observed in the human or mouse transcriptome in normal cells.

We previously proposed that immunostimulatory and proinflammatory properties of highly inflammatory influenza and other RNA viruses derive, in part, from RNA containing CpGs in AU-rich contexts, which are avoided in RNA viruses circulating in humans. Experimental evidence has supported this hypothesis (23, 38, 39). Recently we recast our analysis in the language of statistical physics in a way that is theoretically insightful and computationally efficient (15). In this language, the evolution and optimization of nucleotide sequence motifs are driven by the interplay between selective and entropic forces. The latter randomize motif frequencies in a genome under constraints, whereas the former are largely Darwinian, optimizing for functions enhancing viral replication and spread. However, ncRNAs transcribed mostly in cancerous cells would not be exposed to the same selective and entropic forces as coding RNAs and ncRNAs transcribed in normal cells. Based on motif use patterns, we predicted many ncRNAs may have immunogenic properties, presenting DAMPs.

We focused experimentally on HSATII and murine GSAT, because they are preferentially and highly expressed in carcinogenic processes and exhibit abnormal patterns of motif use. In particular, human HSATII is enriched in CpG motifs in AU-rich contexts avoided in genomes of humans and human-adapted viruses. We demonstrate that their computationally predicted immunogenic properties lead to the induction of inflammatory cytokines in human and murine innate cells (Fig. 3 *A* and *B*). Our observations, together with previous work by Leonova et al. (8), strongly suggest that these endogenous i-ncRNAs are recognized as DAMPs by cellular nucleic acid PRRs.

We identified a key role for MYD88 and UNC93b as regulators of GSAT immunogenicity, but without evidence for the common endosomal nucleic acid sensors typically regulated by UNC93b or associated with the MYD88 adaptor (TLR2, TLR4, TLR7, and TLR9). Our results indicate that in the murine imBM background, there is potent induction of TNF-alpha. Further studies will be required to elucidate whether TLR13, which has been identified in murine cells and recognizes ribosomal bacterial and viral RNA, is involved, or whether there exist intracellular sensors of i-ncRNA associated with MYD88 (40–42), as there are for dsDNA (DHX-9 or DHS-36) (43). Interestingly, we find alignment of GSAT contains a subsequence conserved in immunogenic RNA isolated from bacterial ribosomal RNA, which specifically activates murine TLR13 (41).

Activation of innate immune signaling can contribute to either carcinogenesis or antitumoral immunity. TLR signaling and MYD88 have been associated with tumor development (44). Given that HSATII and GSAT expression has been found to be pervasive in many tumor types and induces responses that differ by species or cell type, the role of i-ncRNA in tumorigenesis is likely dependent on the particular RNA expressed and other properties of the tumor microenvironment. For instance, HSATII activates macrophages and monocytes in our study, suggesting it may be a mechanism for attraction and retention of tumor-associated macrophages. These macrophages have consistently been shown to be a poor prognostic in cancer, leading to increased tumorigenesis, metastasis, and immunoevasion (45). Under this hypothesis, HSATII is used by the tumor to keep macrophages in the tumor microenvironment while driving out T cells. Interestingly, the viral-like behavior of HSATII transcripts is found not only in the immune response to these elements but also in their ability to reverse-transcribe in cancer cells, akin to retroviruses (46).

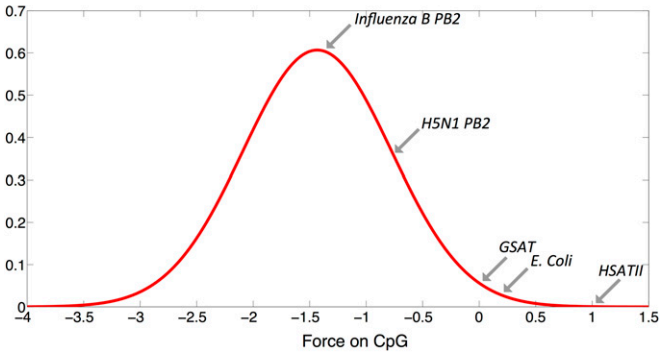


Fig. 5. Motif use in HSATII and GSAT clusters with foreign RNA. A comparison of the forces on CpG dinucleotides is plotted against the distribution of forces on all GENCODE lncRNA relative to a sequences nucleotide bias. The force on CpG dinucleotides for HSATII and GSAT is shown on the distribution, along with the average values for the longest gene (PB2) in human influenza B and avian H5N1 and all *Escherichia coli* coding regions.

The i-ncRNA, not subject to the same forces as ncRNA transcribed in steady state, may retain or evolve to mimic features of foreign RNA, as seen by comparing HSATII and GSAT with typical human ncRNA and foreign genomic material in Fig. 5 (15, 47). Indeed, HSATII and GSAT cluster more closely, in terms of motif use patterns, with bacterial rather than human RNA. Such RNA may have been selected to identify and eliminate cells when their epigenetic state is disrupted. Essentially self-"junk" RNA may have been maintained or may have evolved to mimic non–self-pathogen–associated patterns to create a danger signal. We propose that such a mechanism would be a previously unidentified aspect of "genetic mimicry," where the host is, for all practical purposes, mimicking pathogen-associated nucleic acid patterns. HSATII and GSAT emanate from the pericentromeres, which harbor new repetitive elements with no known function (48). This region, unlike centromeres or regions critical for structure or regulation, may dynamically produce unusual repetitive elements that can adapt to a particular organism's PRRs. Our studies indicate that under the "extraordinary" circumstances where these repetitive elements are expressed, they could play a critical role in the regulation of immune responses against cancer.

Materials and Methods

We consider an RNA sequence of length L, hereafter called S_0, and a motif m [a series of contiguous nucleotides (e.g., CpG)]. Our objective is to define a probabilistic model over the set of the 4^L sequences, $S=(s_1\ s_2 \ldots\ s_i \ldots\ s_L)$, such that the average value of the number, $N_m(S)$, of occurrences of the motif m in S coincides with the number, $N_m(S_0)$, of occurrences of that motif in S_0. To do so, we consider a random-nucleotide model, where nucleotides are independently distributed according to the frequencies $f^0(s)$, with $s=A,C,G,U$, found in S_0. We then introduce the weakest bias that allows us to reproduce $N_m(S_0)$ on average.

The probability of a sequence S in this least-constrained, maximum entropy model is

$$P(S|x,m)=\frac{1}{Z_m(x)}\prod_{i=1}^{L} f^0(s_i)\ exp(x\ N_m(S)), \qquad [1]$$

where

$$Z_m(x) = \sum_{\text{sequences } S} \prod_{i=1}^{L} f^0(s_i)\ exp(x\ N_m(S)) \qquad [2]$$

ensures the probability is correctly normalized. Parameter x, referred to as a selective force (or just force) on the motif m, introduces a statistical bias over P (15). The force quantifies the strength of statistical bias, which may be due to selection on a motif. In the absence of bias ($x = 0$), the probability of S simplifies to the product of its nucleotide frequencies, and the number of motifs is what one would expect in a typical sequence with nucleotide frequencies given by $f^0(s)$. Positive values for x push the distribution toward sequences with $N_m(S)$ larger than what one would expect, whereas negative values for x favor sequences with a smaller $N_m(S)$ than expected.

The value of the force, $x(S_0)$, is computed by maximizing the probability $P(S_0|x,\ m)$ of the sequence S_0 over x. This calculation is equivalent to finding the value of x such that the average number of motifs,

$$N_m^{av}(x) = \sum_{\text{sequences } S} P(S|x,\ m)\ N_m(S) = \frac{\partial \log Z_m}{\partial x}(x), \qquad [3]$$

equals $N_m(S_0)$. By scanning the sequences S_0 in the GENCODE database, we obtain the forces $x(S_0)$ shown in Fig. 2.

The logarithm of the number of sequences having $N_m(S)$ repetitions of m is bounded from above by the entropy of the random-nucleotide model; the equality is reached in the absence of bias only ($x = 0$). The difference between those entropies is the entropy cost corresponding to the constraint on the average number of occurrences of m, and is denoted by σ_m. It is the Legendre transform of $\log Z_m(x)$ (Eqs. **2** and **3**):

$$\sigma_m = x(S_0)\ N_m(S_0) - \log Z_m(x(S_0)). \qquad [4]$$

Efficient computational techniques allow us to calculate the sum over the 4^L sequences in Eq. **2** in a time growing only linearly with L.

Our aim is to find anomalous motif use in a sequence where the number of motif occurrences is different from what is expected by chance in the random-nucleotide model (i.e., associated with a significant nonzero force). We express the likelihood of observing the natural sequence S_0 with a given motif count as

$$P(S^0|m) = \max_x \left[P(S^0|x,\ m)\right] = e^{\sigma_m} \prod_i f^0(s_i^0). \qquad [5]$$

This likelihood is therefore directly related to the entropic cost: The larger the cost, the more likely is the motif to be statistically significant.

ACKNOWLEDGMENTS. We thank Dr. K. Fitzgerald (University of Massachusetts Medical School), Dr. R. Vance (University of California, Berkeley), Dr. G. Barton (University of California, Berkeley), and the Biodefense and Emerging Infections Research Resources Repository [American Type Culture Collection/National Institute of Allergy and Infectious Diseases (NIAID)] for helping us collect murine immortalized macrophages. We also thank Dr. N. Vabret for many helpful discussions and A. Munk for all of his assistance. B.D.G. was supported by NIH [National Cancer Institute (NCI)] Grant 5P01CA087497-13; N.B. was supported by NIH (NIAID) Grants 5R01AI081848-05 and 5R01AI081848-05, NCI Grant 1R01CA180913-01A1, and the Cancer Research Institute; D.T.T. was supported by NIH (NCI) Grant K12CA087723-11A1, Department of Defense (US Army) Grant W81XWH-13-1-0237, and the Burroughs Welcome Fund; and R.M. and S.C. were supported by L'Agence Nationale de la Recherche Grant ANR-13-BS04-0012-01.

1. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
2. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
3. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.
4. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nat Rev Genet* 12(10):671–682.
5. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.
6. Atianand MK, Fitzgerald KA (2013) Molecular basis of DNA recognition in the immune system. *J Immunol* 190(5):1911–1918.
7. Zhang K, et al. (2014) The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene* 547(1):1–9.
8. Leonova KI, et al. (2013) p53 cooperates with DNA methylation and a suicidal interferon response to maintain epigenetic silencing of repeats and noncoding RNAs. *Proc Natl Acad Sci USA* 110(1):E89–E98.
9. Ting DT, et al. (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331(6017):593–596.
10. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. *Science* 293(5532):1068–1070.
11. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nat Rev Cancer* 4(2):143–153.
12. Yi L, Lu C, Hu W, Sun Y, Levine AJ (2012) Multiple roles of p53-related pathways in somatic cell reprogramming and stem cell differentiation. *Cancer Res* 72(21):5635–5645.
13. Zeng M, et al. (2014) MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. *Science* 346(6216):1486–1492.
14. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. *PLoS Pathog* 4(6):e1000079.
15. Greenbaum BD, Cocco S, Levine AJ, Monasson R (2014) Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. *Proc Natl Acad Sci USA* 111(13):5054–5059.
16. Ulitsky I, Bartel DP (2013) lincRNAs: Genomics, evolution, and mechanisms. *Cell* 154(1):26–46.
17. Levine AJ, Greenbaum B (2012) The maintenance of epigenetic states by p53: The guardian of the epigenome. *Oncotarget* 3(12):1503–1504.
18. Coleman JR, et al. (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* 320(5884):1784–1787.
19. Mueller S, et al. (2010) Live attenuated influenza virus vaccines by computer-aided rational design. *Nat Biotechnol* 28(7):723–726.
20. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E (2006) Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J Virol* 80(19):9687–9696.
21. Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* 68(5):2889–2897.
22. Greenbaum BD, Rabadan R, Levine AJ (2009) Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. *PLoS One* 4(6):e5969.
23. Jimenez-Baranda S, et al. (2011) Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. *J Virol* 85(8):3893–3904.
24. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110(1-4):462–467.
25. Xie C, et al. (2014) NONCODEv4: Exploring the world of long non-coding RNA genes. *Nucleic Acids Res* 42(Database issue):D98–D103.
26. Guilliams M, et al. (2014) Dendritic cells, monocytes and macrophages: A unified nomenclature based on ontogeny. *Nat Rev Immunol* 14(8):571–578.
27. Kroemer G, Galluzzi L, Kepp O, Zitvogel L (2013) Immunogenic cell death in cancer therapy. *Annu Rev Immunol* 31:51–72.
28. Sabado RL, Bhardwaj N (2013) Dendritic cell immunotherapy. *Ann N Y Acad Sci* 1284: 31–45.
29. Casrouge A, et al. (2006) Herpes simplex virus encephalitis in human UNC-93B deficiency. *Science* 314(5797):308–312.
30. Lee BL, et al. (2013) UNC93B1 mediates differential trafficking of endosomal TLRs. *eLife* 2:e00291.
31. Tabeta K, et al. (2006) The Unc93b1 mutation 3d disrupts exogenous antigen presentation and signaling via Toll-like receptors 3, 7 and 9. *Nat Immunol* 7(2):156–164.
32. O'Neill LA, Golenbock D, Bowie AG (2013) The history of Toll-like receptors - redefining innate immunity. *Nat Rev Immunol* 13(6):453–460.
33. Broz P, Monack DM (2013) Newly described pattern recognition receptors team up against intracellular pathogens. *Nat Rev Immunol* 13(8):551–565.
34. Gajewski TF, Schreiber H, Fu YX (2013) Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* 14(10):1014–1022.
35. Bogunovic D, et al. (2009) Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc Natl Acad Sci USA* 106(48):20429–20434.
36. Kayagaki N, et al. (2011) Non-canonical inflammasome activation targets caspase-11. *Nature* 479(7371):117–121.
37. Cosset É, et al. (2014) Comprehensive metagenomic analysis of glioblastoma reveals absence of known virus despite antiviral-like type I interferon gene response. *Int J Cancer* 135(6):1381–1389.
38. Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P (2014) The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. *Nucleic Acids Res* 42(7):4527–4545.
39. Vabret N, et al. (2012) The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. *PLoS One* 7(4):e33502.
40. Li XD, Chen ZJ (2012) Sequence specific detection of bacterial 23S ribosomal RNA by TLR13. *eLife* 1:e00102.
41. Oldenburg M, et al. (2012) TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. *Science* 337(6098):1111–1115.
42. Shi Z, et al. (2011) A novel Toll-like receptor that recognizes vesicular stomatitis virus. *J Biol Chem* 286(6):4517–4524.
43. Kim T, et al. (2010) Aspartate-glutamate-alanine-histidine box motif (DEAH)/RNA helicase A helicases sense microbial DNA in human plasmacytoid dendritic cells. *Proc Natl Acad Sci USA* 107(34):15181–15186.
44. Wang JQ, Jeelall YS, Ferguson LL, Horikawa K (2014) Toll-like receptors and cancer: MYD88 mutation and inflammation. *Front Immunol* 5:367.
45. Noy R, Pollard JW (2014) Tumor-associated macrophages: From mechanisms to therapy. *Immunity* 41(1):49–61.
46. Bersani F, et al. (2015) Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proc Natl Acad Sci*, in press.
47. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12(6):996–1006.
48. Maumus F, Quesneville H (2014) Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun* 5:4104.

SUPPLEMENTARY METHODS AND EXPERIMENTS

Design of Experimental Controls. For HSATII and GSAT negative controls were designed in two ways and both negative controls were compared to HSATII and GSAT for all experiments. First, full RNA sequences of both satellites were randomly permuted until scrambled sequences were generated that fell within one half of a standard deviation from the mean value of the strength of statistical bias against CpG and UpA dinucleotides for humans and mice respectively. These sequences are denoted as HSATII-sc and GSAT-sc. In other words these sequences had the same length and nucleotide content as HSATII and GSAT but fell within the inner ellipse in Figures 2a (HSATII-sc) and Figure 2b (GSAT-sc). In addition we checked that in both cases the minimum RNA folding energy was not lowered during the scrambling process so that our permutations did not seem to produce more RNA secondary structure thereby creating the possibility of innate immune stimulation via TLR3. The free energy was calculated using the MATLAB RNAfold routine (1,2). We created endogenous negative controls by searching Repbase for the repetitive elements that fell within one standard deviation of the mean bias against CpG and UpA in humans and mice but were also closest in length to HSATII and GSAT. These were UCON38 for HSATII and RMER16A3 for GSAT.

GSAT RNA Expression Level Detection. GSAT RNA expression levels were investigated by a custom Taqman Assay in normal mouse tissue versus mouse tumor tissue samples (Supplementary Figure 3). The tumor mouse models that were investigated were a model of testicular teratoma (p53-/- 129/Sv^{SL}) and a model of liposarcoma ($p53^{LoxP/LoxP}$;$Pten^{LoxP/LoxP}$). In all instances GSAT levels were increased in the tumor samples as compared to normal samples however to varying degrees. There was no significant difference in GSAT levels between tumors arising in females versus those arising in males in the liposarcoma model. Also there was no difference in GSAT levels in p53-/- 129/Sv^{SL} that developed teratomas at a young age (~1 month old) versus at an older age (~3-4 months old) (3,4).

i-ncRNA generation. Sequences encoding for murine GSAT and human HSATII were generated by custom gene synthesis (Genscript) and cloned into a pCDNA3 backbone (EcoRI/EcoRV) that carries a T7 promoter on the + strand and a SP6 promoter on the – strand (Invitrogen). Sequences encoding for GSAT-sc, HSATII-sc, UCON38 and RMER16A3 were generated as minigenes and sub-cloned in a pIDT-blue backbone with a T7 promoter on the + strand and a T3 promoter on the – strand surrounding the sequence of interest (IDT). To produce high quality RNA, plasmids were digested by the restriction enzymes NotI/NdeI (pCDNA3) and ApaLI (pIDT blue) to isolate the fragment containing the sequence of interest by gel purification (Qiagen). Then the sequences of interest containing the T7 promoter were amplified by PCR (Accuprime-PFX Invitrogen) using the following primer pairs:
pIDT blue - Forward: GCGCGTAATACGACTCACTATAGGCGA;
Reverse: CGCAARRAACCCTCACTAAAGGGAACA) and

pCDNA.3 - Forward: GAAATTAATACGACTCAATAGG;
Reverse:TCTAGCATTTAGGTGACACTATAGAATAG).
PCR products were purified by PCR-Cleanup (Qiagen) and controlled by electrophoresis (0 8% Agarose gel). RNAs were generated by in-vitro transcription using the mMESSAGE mMACHINE T7 ultra kit (Ambion) followed by a capping and short polyA reaction. RNAs were then purified using RNA-cleanup (Qiagen) quantified using a nanodrop and checked by electrophoresis after denaturation at 65 C for 10 minutes (1 5% Agarose gel).

Cell stimulation. MoDCs and imBM were both stimulated by i-ncRNA in the same way. The culturing of these cells is described below. Briefly cells were plated in 96 flat well plates at 200,000 cells per well for primary cells (MoDCs) and 100,000 cells per well for lines (IMBM). i-ncRNA were transfected via liposomes formed using DOTAP (Roche Life Science) at a ratio of 1ug DNA per 6 ul DOTAP diluted in HBS following the user-guide recommendations. The cells were stimulated using 2ug/ml of purified i-ncRNA versus 10ug/ml total RNA. To stimulate the TLR4 pathway we used 100ng/ml Ultrapure LPS (Invivogen) for TLR2: 500ng/ml Pam2CSK4 (Invivogen) for TLR3: 2ug/ml HMW PolyIC (Invivogen) TLR7/8: 1ug/ml CLO97 (Invivogen) and 100 ng/ml R848 (Invivogen) TLR9: CpG B-ODN 1826 3uM or STING CDN 5ug/ml (Aduro).

Cell culture. *Human moDCs:* Human monocyte derived DCs were differentiated as previously described (5); briefly PBMCs were prepared by centrifugation over Ficoll-Hypaque gradients (BioWhittaker) from healthy donor buffy coats (New York Blood Center). Monocytes were isolated from PBMCs by adherence and then treated with 100 U/ml GM-CSF (Leukine Sanofi Oncology) and 300 U/ml IL-4 (RandD) in RPMI plus 5% human AB serum (Gemini Bio Products). Differentiation media was renewed on day 2 and day 4 of culture. Mature moDCs were harvested for use on days 5 to 7. For all experiments harvested DCs were washed and equilibrated in serum-free X-Vivo 15 media (Lonza).
Murine imBMs: Immortalized macrophages were immortalized by infecting bone marrow progenitors with oncogenic v-myc/vraf expressing J2 retrovirus as previously described (6) and differentiated in macrophage differentiated media containing MCSF. ImBM were maintained in 10% FCS PSN DMEM (Gibco). ImBM lines have been kindly provided by several collaborators and also obtained from the BEI resource: ICE (Casp1/Casp11), MAVs, IFN-R, IRF3-7 (Dr. K.Fitzgerald University of Massachusetts), STING and their rescues (Dr. R. Vance University of California Berkeley), Unc93b1 3d/3d (Dr. G. Barton University of Californina Berkeley), TLR 3, 4, 7, 9, 2-9, 2-4, MYD88, TRIF, TRAM, TRIF-TRAM (BEI resource ATCC/NIAID).

Investigation of Type I Interferon Pathway. To characterize whether this pathway could be modulated in our models, we evaluated production of type I interferon in response to stimulation by our i-ncRNA using human and murine interferon stimulated response element (ISRE) reporter cell lines, and monitored transcriptome regulation of a panel of immune genes related to the interferon pathway. Whereas the effect on the

inflammatory response is significant in terms of TNFalpha, IL-6, or IL-12 production, the effect on the type I interferon pathway was less prominent.

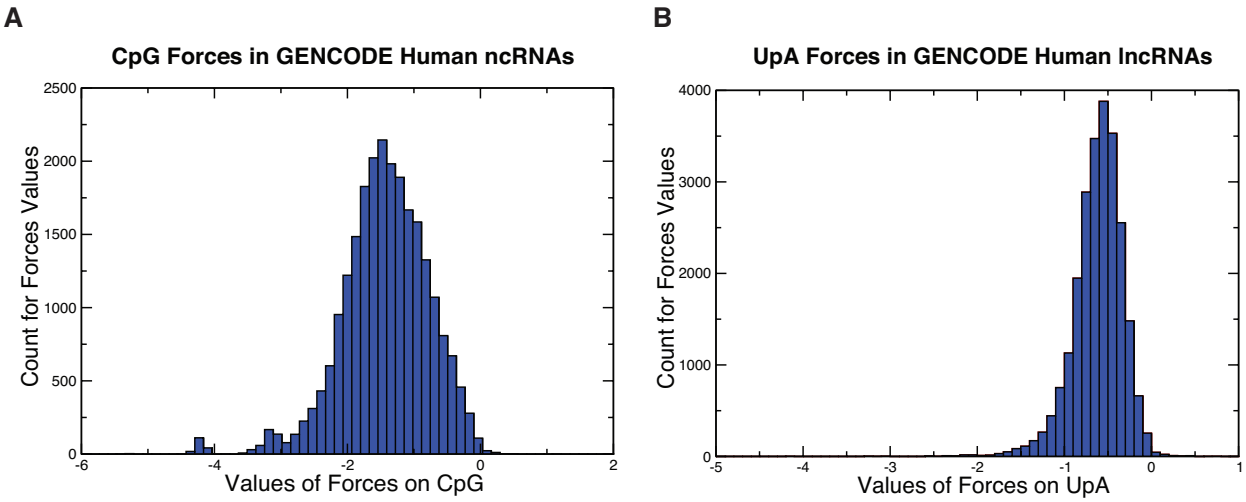
Additional Pathways Investigated. TLR2 or TLR4 were not required, indicating the observed effect was independent of contamination from bacterial products such as lipoproteins and endotoxins (Supplementary Figure 8). TRIF, TRIF/TRAM, and IRF3/IRF7, which participate down stream in the signaling of TLR3, TLR4, and TLR7, were also not obligatory (Supplementary Figure 5). We did not identify a role for candidate molecules for sensing murine GSAT, such sensors related to cGAS-STING signaling or DEAD box RNA helicases such as RIG-I and MDA5 (7-10). Inflammatory responses to GSAT did not depend upon the stimulator of interferon genes (STING), which induces type I interferon production when cells are infected with intracellular pathogens. RIG-I (retinoic acid-inducible gene 1) is a dsRNA helicase enzyme that senses RNA viruses through activation of the mitochondrial antiviral-signaling protein (MAVS) (11-13). MAVS deficient imBMs failed to respond to GSAT stimulation ruling out a contribution of RIG-I in our i-ncRNA signaling (Supplementary Figure 7B). Finally we ruled out a role for inflammasome related pathways using ICE-KO imBM that are essentially a knockout for Caspase 1 and which carry an inactive mutation for Caspase 11.

SUPPLEMENTARY REFERENCES

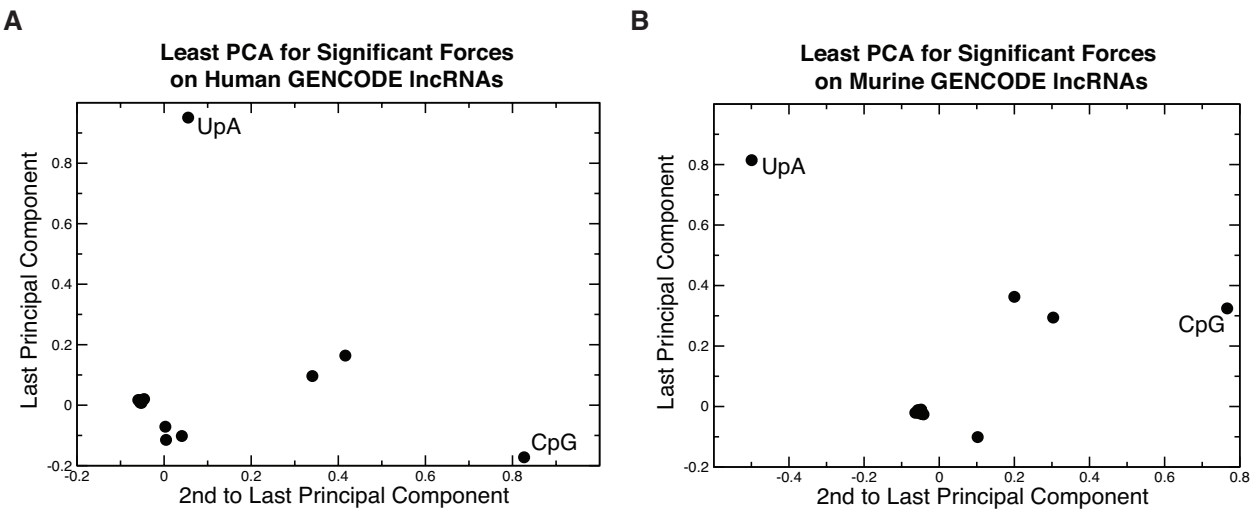
1. Matthews D, Sabina J, Zuker M, Turner D (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288:911–940.
2. Wuchty S, Fontana W, Hofacker I, Schuster P (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* 49:145-165.
3. Harvey M, McArthur MJ, Montgomery CA, Bradley A, Donehower LA (1993) Genetic background alters the spectrum of tumors that develop in p53-deficient mice. *The FASEB Journal* 7:938-943.
4. Muller AJ, Teresky AK, Levine AJ (2000) A male germ cell tumor-susceptibility determining locus pgct1 identified on murine chromosome 13. *Proc Natl Acad Sci* 97:8421-8426.
5. Frleta D, et al. (2012) HIV-1 infection–induced apoptotic microparticles inhibit human DCs via CD44. *J Clinical Inves* 122:4685.
6. Blasi E, et al. (1985) Selective immortalization of murine macrophages from fresh bone marrow by a raf/myc recombinant murine retrovirus. *Nature* 318:667-670.
7. Atianand MK, Fitzgerald KA (2013) Molecular basis of DNA recognition in the immune system. *J Immunol* 190:1911-1918.
8. Lee BL, et al. (2013) UNC93B1 mediates differential trafficking of endosomal TLRs. *eLife* 2:e00291.
9. Burdette DL, Vance RE (2013) STING and the innate immune response to nucleic acids in the cytosol. *Nature Immunol* 14:19-26.
10. Vanaja SK, Rathinam VA, Fitzgerald KA (2015) Mechanisms of inflammasome

activation: recent advances and novel insights. *Trends Cell Biol*, in press.
11. Zeng M, et al. (2014) MAVS cGAS and endogenous retroviruses in T-independent B cell responses. *Science* 346:1486-1492.
12. Broz P, Monack DM (2013) Newly described pattern recognition receptors team up against intracellular pathogens. *Nature Rev Immunol* 13:551-565.
13. Gajewski TF, Schreiber H, Fu YX (2013) Innate and adaptive immune cells in the tumor microenvironment. *Nature Immunol* 14:1014-1022.

SUPPLEMENTARY FIGURES

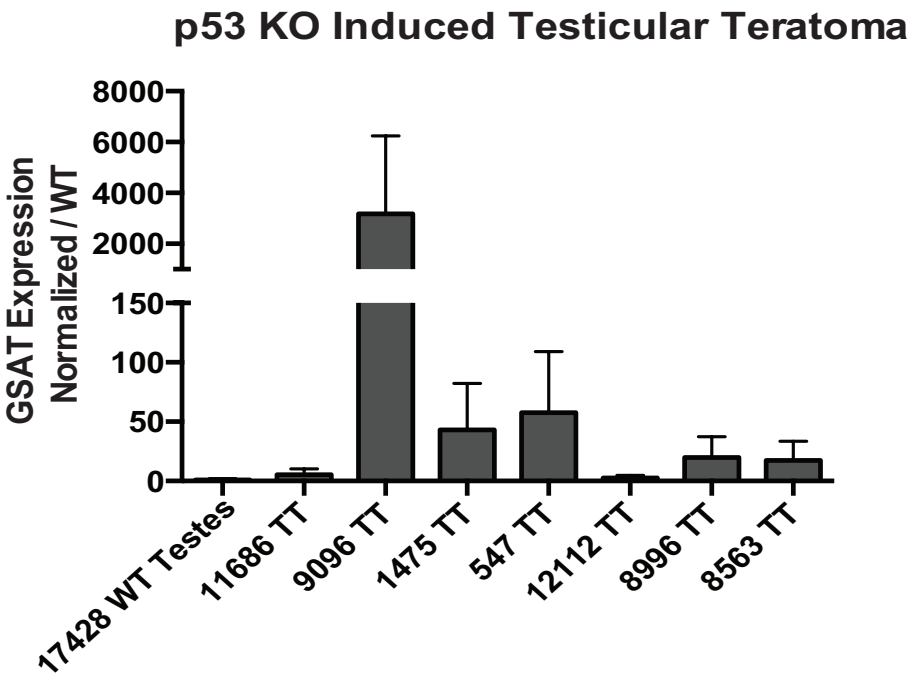


Supplementary Figure 1. CpG and UpA Are Generally Under-represented in ncRNA. Histogram of forces (strength of statistical bias) on (A) CpG and (B) UpA for lncRNA from the GENCODE Human transcript database. These forces are consistent with those observed in mice and those from coding regions.

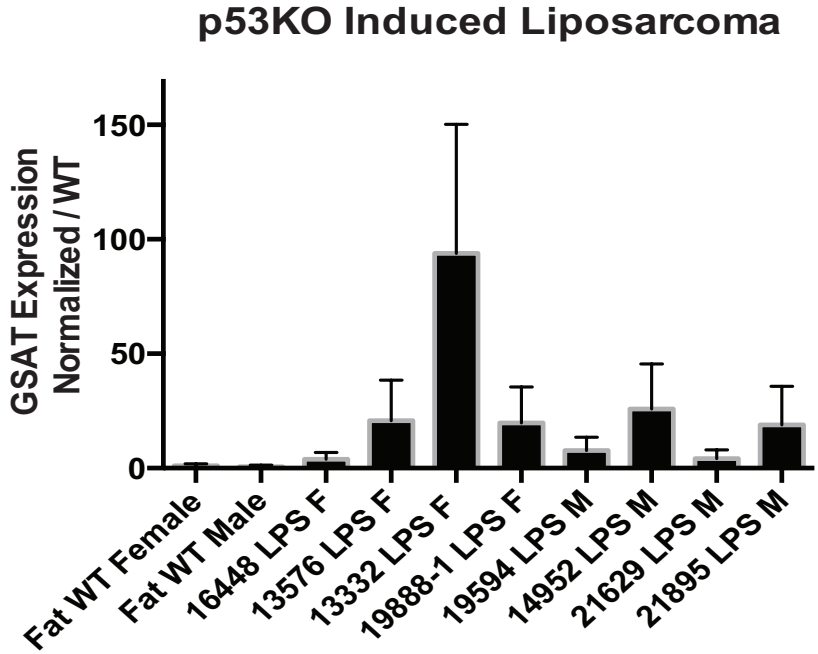


Supplementary Figure 2. Forces on CpG and UpA Dinucleotides Are Independent. Least principal components for all significant forces on motifs for (A) human and (B) mouse GENCODE ncRNA. In both cases CpG and UpA dominantly project onto the two least axes of variation.

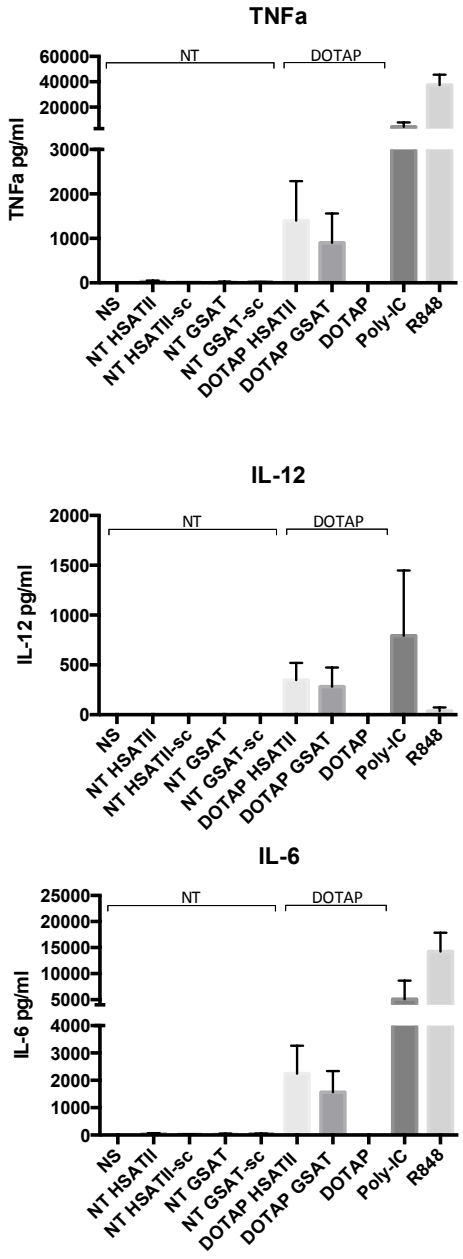
A



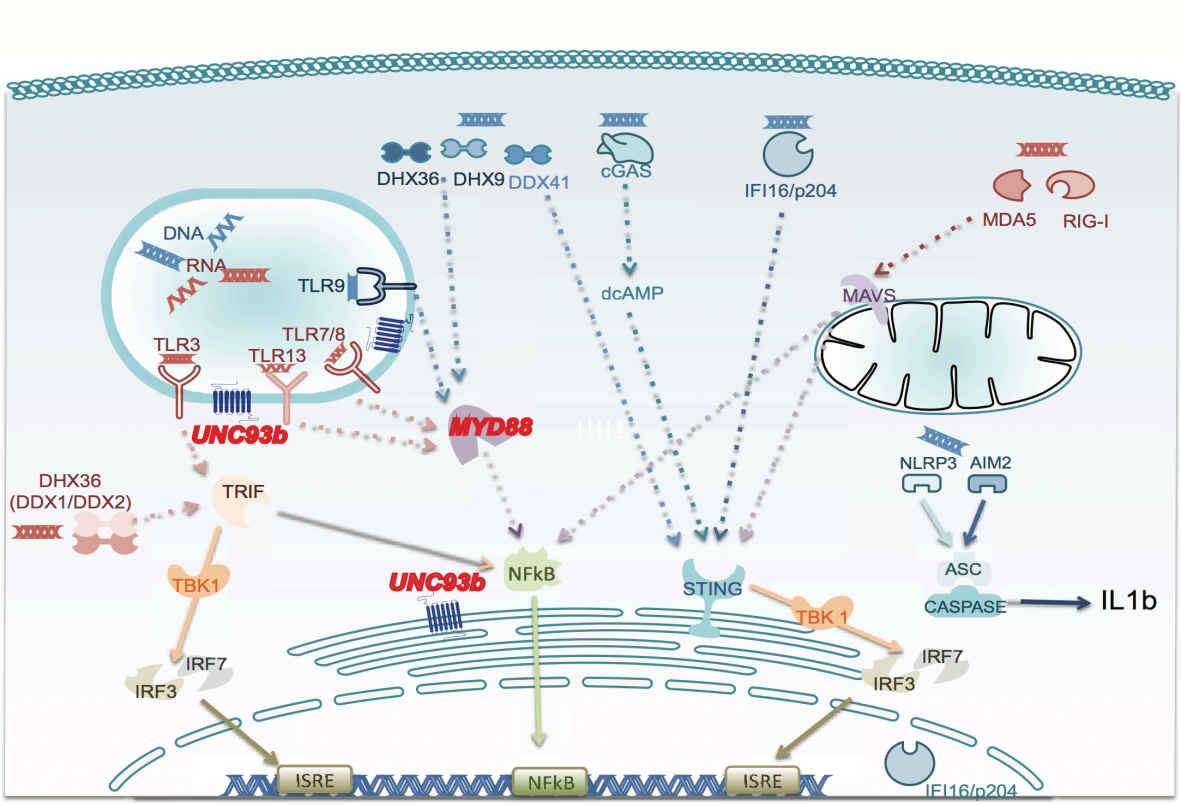
B



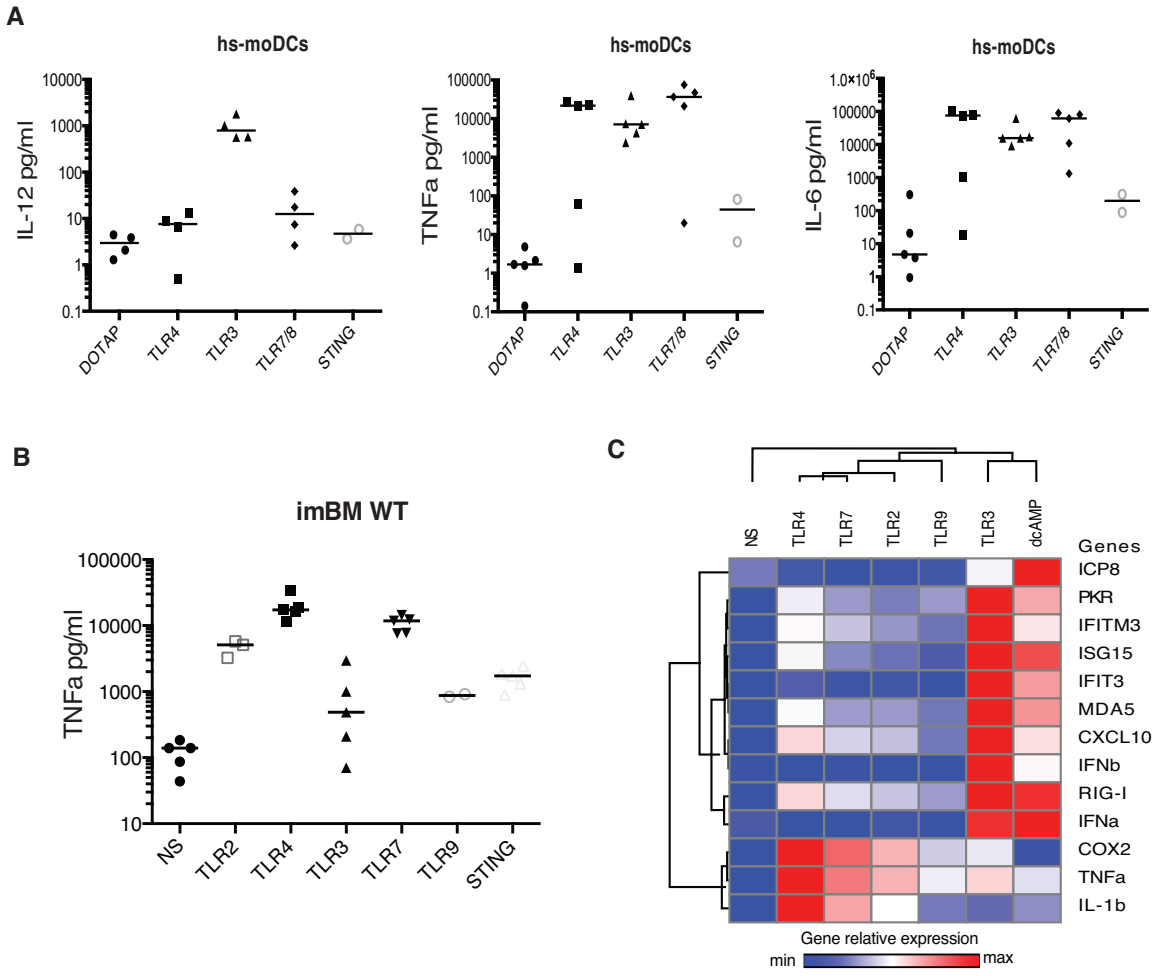
Supplementary Figure 3. GSAT is Expressed in Mouse Testicular Teratoma and Liposarcoma. Study of the relative levels of expression of GSAT RNA by a custom Taqman Assay in normal murine tissue versus murine tumor tissue samples. The tumor mouse models investigated were testicular (A) teratoma and (B) liposarcoma induced tumor in p53KO background. In all instances, GSAT levels were increased in the tumor samples as compared to normal samples, to varying degrees.



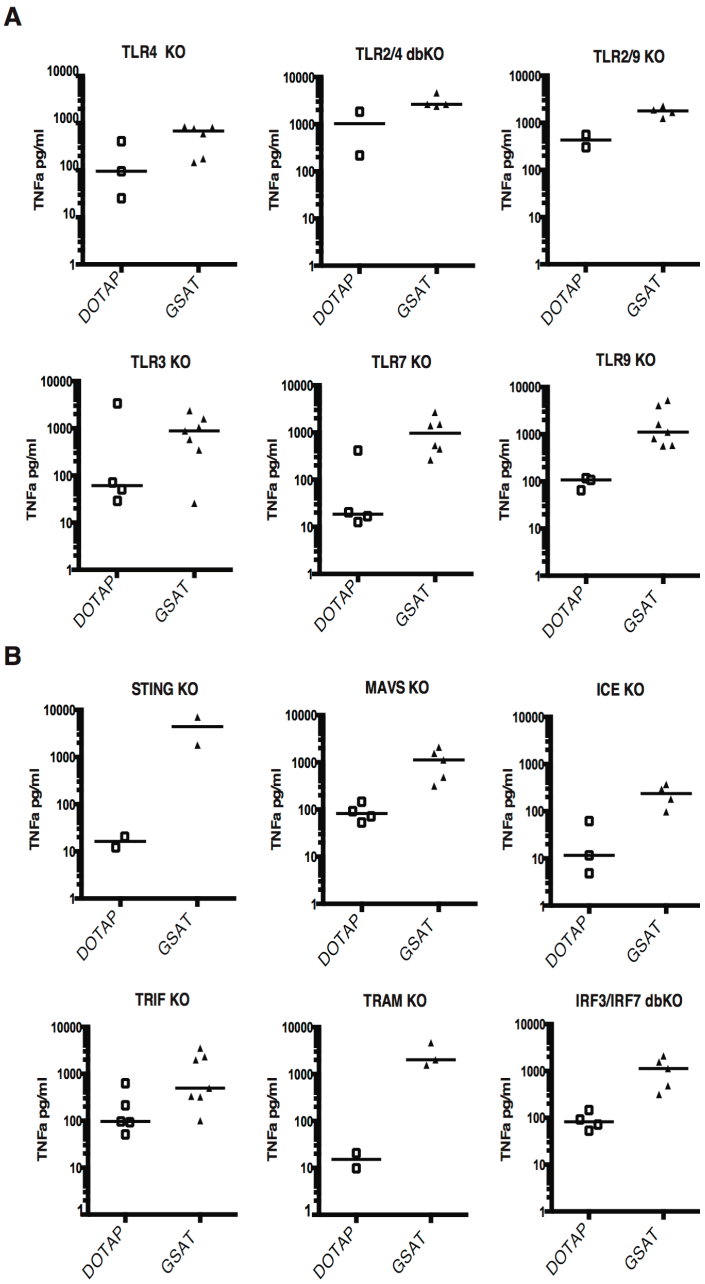
Supplementary Figure 4. NcRNA Require Transfection to Induce Cellular Innate Immune Responses. 2ug /ml of the various ncRNA (HSAT II, HSAT II-sc; GSAT; GSAT-sc) were used to stimulate human DCs in 96 well plates with (DOTAP) or without (NT) the use of DOTAP as a gentle liposomal transfection reagent. In absence of transfection reagent the ncRNA were not sensed by the DCs whereas transfected immunogenic ncRNA HSAT II and GSAT, in addition to Poly-IC and R848, were properly sensed and induced a cellular inflammatory response in (A) TNFalpha, (B) IL-12, and (C) IL-6.



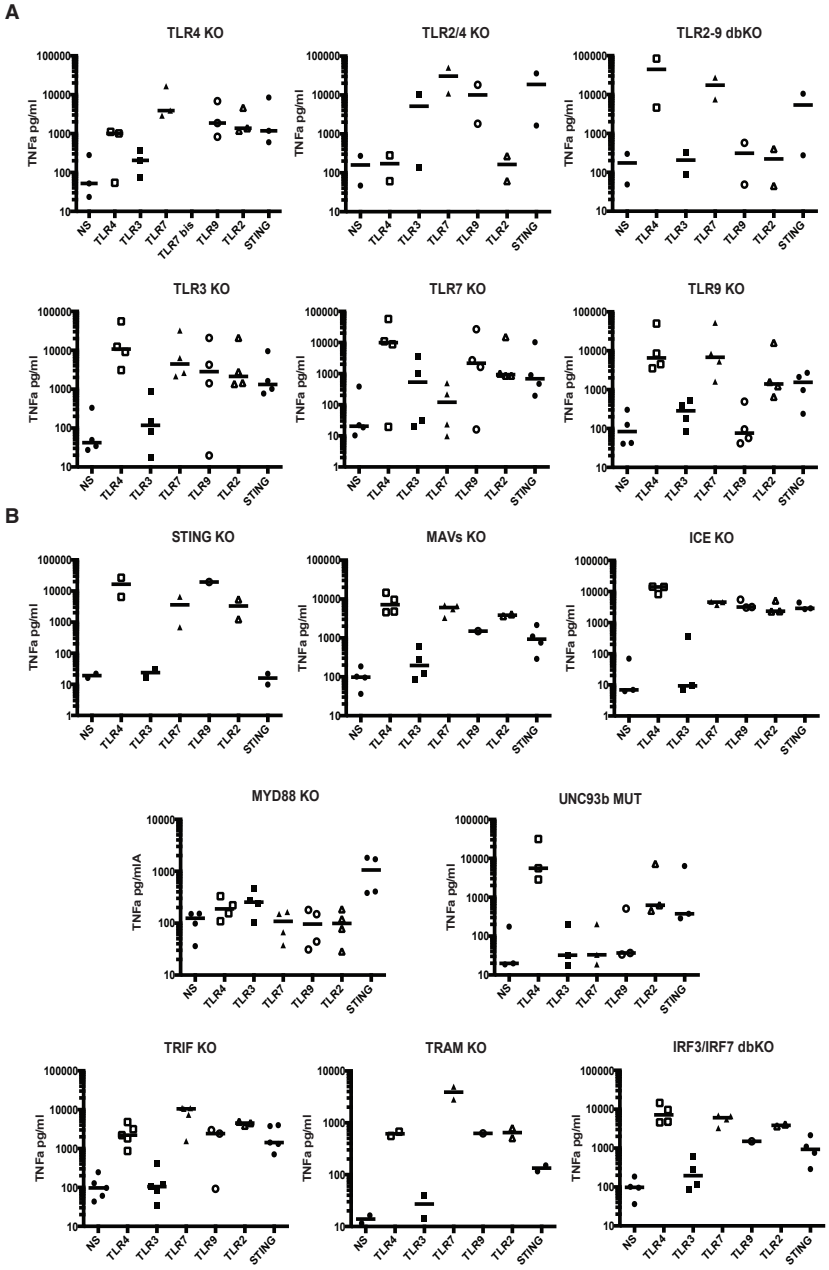
Supplementary Figure 5. Innate Immune Sensing of Nucleic Acids. Summary of the innate immune pathways involved in the sensing of nucleic acids which were investigated in this work. MYD88 and UNC93b, highlighted in red, were directly implicated in i-ncRNA sensing.



Supplementary Figure 6. Human moDCs and Mouse imBM Cells Respond to Common PAMPs and DAMPs. Quantification of inflammatory cytokine production in human moDCs (A) and in murine imBM (B) upon stimulation with common PAMPs or DAMPs known to activate PRR innate immune pathways, which are listed in the Materials and Methods. Each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median. (C) The inflammatory response related to type I IFN pathway induction in imBM upon stimulation of the PRR related innate immune pathways has been analyzed by qRT-PCR. The heat-map represents the log of the relative expression of each gene based on relative quantification analysis using the ddCT bi-dimensional normalization method (house keeping genes and non-stimulated cells).



Supplementary Figure 7. Genetic Screen of Innate Immune Pathways Related to i-ncRNA Function in Murine imBM. (A) imBM cells of different knockout genotypes related to TLR PRRs (TLR2-4 dbKO, TLR3 KO, TLR4 KO, TLR7 KO, TLR9 KO). (B) imBM cells of different knockout genotypes related to STING, inflammasome, and MAV dependent helicases pathways (STING KO, MAV KO, ICE KO); and common innate immune signaling (TRIF KO, TRAM KO, IRF3/IRF7 dbKO). Cells have been stimulated by liposomal transfection of the murine i-ncRNA (GSAT). The TNFa production in the supernatant has been quantified and each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median.



Supplementary Figure 8. Stimulation of KO and Mutant imBM with Common PAMPs and DAMPs. Quantification of inflammatory cytokine production in PRR KO imBM (A) and innate immune signaling related KO and mutant (B) upon stimulation with common PAMPs or DAMPs known to activate PRR innate immune pathways. Each point represents the mean value of the experimental replicates for each individual condition; the bar represents the median.

SUPPLEMENTARY TABLES

	Human	Mouse
CG	-1.419	-1.375
UA	-0.604	-0.548
ACG	-1.7586	-1.6216
CAG	0.5534	0.5612
CCG	-1.5095	-1.3287
CGA	-1.8995	-1.7082
CGC	-1.7304	-1.5525
CGG	-1.511	-1.2629
CGU	-1.7833	-1.6463
CUG	0.669	0.6748
GCG	-1.748	-1.5592
GUA	-0.8632	-0.7451
UAC	-0.7368	-0.6298
UAG	-0.733	-0.592
UCG	-1.9391	-1.7049

Supplementary Table 1. Average Forces on Motifs are Similar Between Humans and Mice. Average force on a given motif in the Human and Mouse GENCODE dataset, for lncRNAs with length greater than 500 nucleotides. The forces are listed for the significant motifs in humans. The force is a measure of the strength of statistical bias to enhance or suppress a motif versus what is expected from that sequences nucleotide content.

ncRNA	Class	Level of Conservation	CpG Force
MER123	DNA_transposon	Amniota	1.1039
HSATII	SAT	Primates	1.036
UCON21	Transposable_Element	Amniota	0.9465
MER6B	Mariner/Tc1	Homo_spaiens	0.923
Eulor1	Transposable_Element	Amniota	0.8481
Eulor5B	Transposable_Element	Tetrapoda	0.8474
Eulor2C	Transposable_Element	Amniota	0.7676
Eulor6A	Transposable_Element	Tetrapoda	0.7466
MER131	SINE	Amniota	0.6223
Eulor4	Transposable_Element	Tetrapoda	0.6067
Eulor10	Transposable_Element	Amniota	0.6064
MER6C	Mariner/Tc1	Eutheria	0.5667
Eulor12	Transposable_Element	Amniota	0.5295
MER5C1	hAT	Eutheria	0.4582
MER47B	Mariner/Tc1	Eutheria	0.4518
UCON39	DNA_transposon	Mammalia	0.4443
UCON16	Transposable_Element	Amniota	0.4436
Tigger3d	Mariner/Tc1	Primates	0.4374
TIGGER5A	Mariner/Tc1	Eutheria	0.4212
MER75	DNA_transposon	Homo_spaiens	0.4134
Tigger4a	Mariner/Tc1	Primates	0.3815
npiggy2_Mm	piggyBac	Microcebus_murinus	0.3725
MER58B	hAT	Eutheria	0.3657
Eulor6C	Transposable_Element	Tetrapoda	0.3571
Eulor11	Transposable_Element	Amniota	0.3561
UCON15	Transposable_Element	Amniota	0.356
Tigger2b_Pri	Mariner/Tc1	Primates	0.3548
MER44B	Mariner/Tc1	Homo_spaiens	0.3536
SUBTEL_sat	Satellite	Primates	0.3527
Eulor9A	Transposable_Element	Amniota	0.3465
MER44C	Mariner/Tc1	Homo_spaiens	0.3439
Eulor8	Transposable_Element	Amniota	0.3416
MER44D	Mariner/Tc1	Eutheria	0.3211
npiggy1_Mm	piggyback	Microcebus_murinus	0.3131
UCON26	Transposable_Element	Amniota	0.2985
MER127	Mariner/Tc1	Amniota	0.2984
MER97d	hAT	Eutheria	0.2939
Eulor6D	Transposable_Element	Tetrapoda	0.2866
Eulor2B	Transposable_Element	Amniota	0.2852
MER119	hAT	Homo_spaiens	0.2794

MER134	Transposable_Element	Amniota	0.2786
Eulor9C	Transposable_Element	Amniota	0.2751
MER8	Mariner/Tc1	Homo_spaiens	0.2669
Ricksha_a	MuDR	Eutheria	0.2607
MER129	SINE	Amniota	0.2444
MacERV6_LTR3	ERV3	Cercopithecidae	0.2404
MER57B2	ERV1	Homo_spaiens	0.2403
HSMAR1	Mariner/Tc1	Homo_spaiens	0.2397
Eulor12_CM	Transposable_Element	Amniota	0.2269
MERX	Mariner/Tc1	Eutheria	0.2207
Tigger12A	Mariner/Tc1	Mammalia	0.217
MER58A	hAT	Eutheria	0.2006

Supplementary Table 2. Many Repetitive Elements Have High CpG Forces.
Listed above are the repetitive elements from Repbase with a significantly high CpG force. These elements are typically not found to be expressed in normal tissue, yet some may be expressed in cancer cells and cell lines.